HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto Fakultet/Sektion – Faculty	Laitos/Institution– Department
Maatalous-metsätieteellinen	Taloustieteenlaitos

Tekijä/Författare – Author

Jussi Lehtonen

Työn nimi / Arbetets titel – Title

Measuring the impact of substitute sites on the recreational value of the Baltic Sea using the single-site travel cost method

Oppiaine /Läroämne – Subject

Ympäristö- ja luonnonvaraekonomia

Työn laji/Arbetets art – Level	Aika/Datum – Month and year	Sivumäärä/ Sidoantal – Number of pages
pro-gradu tutkielma	05/2019	68

Tiivistelmä/Referat – Abstract

This thesis has three objectives, namely to estimate the recreational value of a Baltic Sea visit, to assess the impact of substitute sites on the value of a recreational visit, and to assess which factors determine whether a visitor has a substitute site for the Baltic Sea if recreation at the preferred site is not possible. No previous Finnish studies have been conducted on this topic, even though the impact of substitute sites is important in the valuation of a recreational site.

This thesis uses the single-site travel cost method to assess the recreational value of the Baltic Sea. The travel cost method is the most frequently used valuation method to assign a monetary value to the recreational services of an environmental resource. To answer the first objective, this thesis employ the single-site travel cost method using the negative binomial model. To answer the second objective, substitutes are entered into the single-site travel cost model as dummy variable and travel cost interaction variable. According to the results of this thesis, the recreational value of a visit to the Baltic Sea is worth between 180 € and 420 €. As expected, substitutes decrease the demand for recreational sites, but, contrary to expectations, the overall value of a single visit to the Baltic Sea is higher among respondents who have substitutes compared to respondents who do not have substitutes. To answer the third objective, the logistic regression model is employed. Based on the results of the model, whether an individual has substitute sites depends on demographics, the number of recreational activities at the Baltic Sea, the timing of the visit and the perceived number of facilities at the Baltic Sea.

Tutkielmalla on kolme tavoitetta. Ensimmäinen tavoite on määrittää Itämeren virkistysarvo, toinen tavoite on määrittää substituuttien vaikutus virkistysarvoon ja kolmas tavoite on selvittää, mitkä tekijät vaikuttavat siihen, onko virkistyjällä substituuttikohdetta Itämeren virkistyspaikalleen. Huolimatta siitä, että substituuttien vaikutus virkistyskohteen arvoon on tärkeä, Suomessa ei ole tehty aikaisempia tutkimuksia aiheesta.

Tässä tutkielmassa Itämeren virkistysarvo määritetään yhden kohteen matkakustannus menetelmää hyödyntäen. Matkakustannus menetelmä on useimmin käytetty arvottamismenetelmä virkistysarvon määrittämiseen. Ensimmäisen tavoitteen toteuttamiseksi on tässä tutkielmassa yhden kohteen matkakustannusmalli estimoitu negatiivisella binomimallilla. Toista tavoitetta varten substituutit lisätään matkakustannusmalliin dummy-muuttujana ja matkakustannusinteraktio -muuttujana. Tutkielman tuloksien mukaan Itämeren virkistyskäynnin arvo on 180€ ja 420€ välillä. Odotetusti substituutit vähentävät virkistyskäyntien kysyntää, mutta vastoin odotuksia Itämeren virkistyskäynnin arvo on suurempi niiden vastaajien keskuudessa, joilla on substituutti verrattuna niihin, joilla ei ole substituuttia Itämeren virkistyskäynnille. Kolmatta tavoitetta varten estimoitiin logistinen regressio. Tulosten perusteella se, onko vastaajalla substituuttia Itämeren virkistyskäynnille riippuu demografisista tekijöistä, kuinka moneen virkistysaktiviteettiin yksilö osallistuu Itämerellä, vierailun ajankohdasta ja Itämeren virkistyskohteen kunnosta.

Avainsanat – Nyckelord – Keywords

Matkakustannusmenetelmä, Itämeri, ympäristön arvottaminen

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

Measuring the impact of substitute sites on the recreational value of the Baltic Sea using the single-site travel cost method

Jussi Lehtonen
Helsingin yliopisto
Taloustieteen laitos
Ympäristö- ja luonnonvaraekonomia
Pro gradu -tutkielma
Toukokuu 2019

Contents

1. Introduction

According to the Baltic Marine Environment Protection Commission (HELCOM, 2016), the Baltic Sea is one of the largest brackish water areas of the world, affecting the lives of 85 million people living in its drainage area. The Baltic Sea offers a variety of goods and services, generally known as ecosystem services. One of the main ecosystem services of the Baltic Sea are its recreational services, for example fishing or walking at the coast. Other services include renewable energy generation and transportation. However, the impaired condition of the Baltic Sea due to eutrophication is well known and has drawn the attention of the residents, the media and policymakers. The poor ecological state of the Baltic Sea is affecting the ecosystem services, thus reducing the welfare of the people. The main objective of this thesis is to estimate the recreational value of the Baltic Sea catchment area for three countries, namely Latvia, Germany and Finland.

Most of the ecosystem services, including recreational services, are characterised as public goods, externalities or common goods, implying a lack of markets for them and thus a lack of prices. The true social value of the services has to be elicited using special measures because market forces fail to reveal the true price and allocation of these resources that do not have markets. There are at least three justifications for the monetary valuation of non-market goods, namely the environmental valuation: 1) the valuation highlights the scarcity of environmental resources and services; 2) policy actions and management decisions influencing environmental resources are more comprehensively evaluated; and 3) the valuation offers a more complete picture of a region's economic performance.

In terms of the environmental valuation, natures benefits are classified under the concept of total economic value, which is divided into two categories, namely use and non-use values. Three different use values are identified: 1) direct use of the environmental resource, for example recreation or timber harvest; 2) indirect use of the resource's ecosystem function, for example timber production; and 3) option values referring to the possible future use of the resource. Two non-use values are generally attached to natural resources: 1) the bequest value, reflecting an individual's willingness to ensure the use of the resource for future generations; and 2) the existence value, referring to the benefit of simply knowing a resource exists.

There are two types of valuation methods, based on how the value of the ecosystem service is determined. Revealed preference method is based on people's actual behavior: people reveal the extent to which they value goods or services by choices. Another method is based on people's stated preferences: people assign a value in hypothetical referendums to changes in natural resources or state how their behaviour could alter if the conditions of the natural resource changed. The valuation method that is used depends on the valuation task at hand.

There are four environmental valuation methods: travel cost, hedonic valuation, contingent valuation and choice experiment. The former two are based on revealed preferences and the latter two on stated preferences. The revealed preference methods are only suitable to estimate use values compared to the stated preference method which capture both non-use and use values.

Three features dictate recreational services at the Baltic Sea: 1) the economic value depends on the environmental condition of the site, for example the amount of fish; 2) access to the site is possible to everyone; and 3) recreational services are not traded at free markets. Thus, there is no market price on the recreational services compared to, for example, transportation services. Environmental valuation methods are needed to elicit the monetary value of the recreational services. Valuing recreational services can be done by using contingent valuation, choice experiments or the travel cost method. The most frequently used method to value recreational services in environmental valuation is the travel cost method.

Between 1995 and 2015, the recreational services of the Baltic Sea were evaluated 29 times (Sagebiel, Schwartz, Rhozyel, Rajmis & Hirschfeld, 2016). Few revealed preference studies exist. Sandström (1996) and Soutukorva (2005) used the random utility maximization travel cost method to value the benefits of Swedish coastal recreation if the water quality would improve due to nutrient reduction. Vesterinen, Pouta, Huhtala and Neuvonen (2010) evaluated the recreational benefits of improving Finland's water systems, including the coast, using the travel cost method. Czajkowski et al. (2015) valued the recreational benefits of all nine coastal countries of the Baltic Sea using the individual travel cost method.

The monetary value elicited by using the travel cost method does include uncertainties arising from the method itself. Accounting for impact of substitutes on the value of recreational services is one of the main issues of the travel cost method and also generally in the valuation of recreational services. particular, substitutes cause problems in valuation of water recreation. The second objective of this thesis is to determine the impact of substitute sites on the monetary value of the Baltic Sea.

The problem of accounting for the impact of substitutes stems from the simple fact that the number of possible substitute sites for water recreational activities is immense, thus limiting the possibilities for including substitute effects in a valuation analysis. In Finnish recreation valuation, substitutes are usually completely omitted from the analysis, based on the fact that "everyman's rights" reinforce two fundamental issues related to substitutes for recreation: 1) the determination of substitutes is difficult; and 2) the number of possible substitutes increases significantly (Huhtala & Lankia, 2012; Lankia et al., 2017). Because of everyman's rights, a clear picture of substitute sites and the impact of

substitute sites on the welfare of recreation is missing. The usual approach of omitting substitutes from the analysis probably inflates the valuation estimates.

A study by Vesterinen et al. (2010), focusing specifically on Baltic Sea recreation, assumed that going to a summer cottage is a good substitute for water recreationists. Random utility models inherently account for substitutes for recreation, but limited attention has been devoted to them in Swedish studies (Sandström, 1996; Soutukorva, 2005). An international study by Czaikowski et al. (2015) completely omitted substitutes.

In this thesis approaches are used to assess the impact of substitute sites on the value of recreation at the Baltic Sea. We employed substitute site dummy variable to assess the overall impact of substitutes on the number of trips taken to the Baltic Sea and travel cost substitute interaction term to assess the sensitivity of the number of trips taken by respondents who have substitute sites on increasing or decreasing travel costs. The third objective of this thesis is to offer insight into what kinds of sites are perceived as substitutes for the Baltic Sea and which visitors have substitute sites for the Baltic Sea using logistic regression.

The research is based on data collected by the BalticApp-project, from 2017 to 2018, from the three countries of Finland, Latvia and Germany. The project's main objective was to identify strategies to ensure the supply of ecosystem services, including recreational services provided by the Baltic Sea in the future.

The results indicate that the annual welfare of the recreational services provided by the Baltic Sea is approximately 3,000€. Substitute sites decrease the number of recreational trips taken to the Baltic Sea, but individuals who have substitute sites are less sensitive to an increase in the travel costs, meaning that even with higher travel costs they take more trips to the Baltic Sea compared to individuals who do not have substitutes. The implication is that respondents who have substitutes are already more committed to recreation at the Baltic Sea. Furthermore, individuals with higher odds of having substitute recreational sites visit the Baltic Sea more often, engage in more than one recreational activity, have a higher education, perceive that their most visited site at the Baltic Sea has many facilities and are between the ages of 30 and 64 years.

The structure of this thesis is as follows. Following the first introductory chapter, Chapter 2 presents a literary review on how recreational substitutes are handled in environmental valuation and how substitute sites impact the monetary value of recreational services. Chapter 3 outlines the data used to conduct the analysis. The fourth chapter introduces the methodology used, namely the travel cost method and logistic regression. Chapter 5 presents the actual travel cost analysis of the Baltic Sea and

the monetary value of a visit to the Baltic Sea. The sixth chapter comprises the discussion and the final chapter presents the conclusions.

2. Literary review: substitution in environmental valuation and social psychology

2.1 Economic theory behind value measure in travel cost model and substitutes

The welfare measures used in environmental valuation is closely linked to neoclassical consumer theory. The theory behind determining value is presented from the perspective of the travel cost method. The essence of consumer theory is that individuals make choices that maximize their utility. However, the utility of an individual cannot be directly observed, but the assumption is that consumers' behavior relies on preferences which reflect individual utility. Individual utility is described in terms of the utility function that describes individuals' preferences for bundles of various goods, where the most preferred bundle of goods yields the maximum utility to the individual (Varian, 2003, pp. 55-58).

The individual utility function is based on a few basic assumptions: 1) the greater the quantity in a bundle of goods, the higher the utility; 2) substitutability exists among the goods in the bundle, meaning that it is possible to substitute one good for another without losing welfare. (Freeman et al., 2013) present the utility function where x is the quantity of market goods, q is the quantity of environmental goods and t is the time spent in activities that yield utility. The difference between *x* and q is that the quantity of *q* is given in the function.

$$U = U(x, q, t)$$

The utility function can be maximized by the following assumptions: 1) prices exist for the goods, in the case of the travel cost model, the assumption is that the price for the environmental good is the cost to travel to the site; and 2) the individual chooses the quantity of goods that maximises the utility. There exists a constraint on individual consumption, which is the amount of disposable income. The solution to the problem results in ordinary Marshallian demand functions. Where $\mathrm{x_i}$ is the quantity of the goods, P is the price of the goods and M is the budget constraint (disposable income).

$$x_i = x_i(P, M)$$

In the travel cost model, welfare is measured as a consumer surplus, which is the area under the Marshallian demand curve, but above the price. Other measures used to calculate environmental valuations are the compensating variation and the equivalent variation; both these measures fall

outside the scope of this thesis (see for example Freeman et al., 2013, pp. 47-59). The impact of substitute sites on the welfare of recreation can be measured as change in the consumer surplus.

Substitute goods or services are alternatives for each other. The substitutability of goods depends on the perceived similarity of the goods. The higher the substitutability, the greater the difference in demand will be. Formally, two goods (A and B) are substitutes if a price increase in good A increases the demand for good B, replacing the consumption of good A with consumption of good B.

In economic terms, two goods are perfect substitutes if they are completely substitutable. In other words, perfect substitutes have a constant marginal rate of substitution (MRS). The MRS is the rate at which a consumer is willing to give up one good in exchange for another good and maintain the same level of utility. Imperfect substitutes have a lesser substitutability level.

In recreation, different sites or recreational activities can serve as substitutes for one another. Our interest here is in the impact of a substitute site on the demand for recreation at the Baltic Sea. The existence of a substitute should lower the demand for recreation as the price of recreation increases. The price increase for recreation means a change in the travel costs to the site under evaluation. The impact of the price change on the good consumed is divided into two effects, the income effect and the substitution effect. The income effect refers to lower or higher consumption of the good and related goods, depending on the price change. The substitution effect means a change in the consumption of the good affected by the price change: if the price of the good increases, the consumption changes to a substitute good if such a substitute good exists. Measuring the substitution effect or the income effect is not that straightforward with regard to environmental goods and services, because of a lack of market prices. In economics, the price of non-market goods is generally called “the shadow price.”

There are two types of substitution effects that should be accounted for in the valuation of recreation sites: 1) The availability of cross-marginal effects, that is, the change in availability of another site in the individual choice set of recreation sites that impacts on the demand for the site under valuation; and 2) the attribute cross-marginal effects, that is, the change in the alternative sites attribute (water quality, new dock, catch rate, etc.) which impacts on the demand for the site under valuation (Schaafsma, 2011, p. 70). For example, in the first case, if the individual acquires information about a good fishing site and access to the new site requires less money than access to the preferred fishing site, the demand for the preferred site decreases. In the second case, if a new road is built to another site, because of the road the shadow price of that site decreases and the demand for the site under valuation decreases.

In this thesis, the assumption is that recreation at the Baltic Sea would not be possible and consumption of recreation would shift to a substitute site if a substitute site exists. The individual's choice set would change, leading to a situation where the shadow price of access to the site under valuation would increase, the demand for recreation at the substitute site would increase and the demand for recreation at the Baltic Sea would decrease. The travel cost model would produce a conventional downward-sloping demand curve for recreation at the Baltic Sea and the value of recreation at the Baltic Sea would be measured as a consumer surplus, as the demand for recreation at the Baltic Sea would decrease and the welfare of recreation at the Baltic Sea would decrease.

2.2 Summaries from previous findings based on meta-analysis

A meta-analysis combines data from multiple studies. In environmental valuation, a meta-analysis is often used to synthesise the benefit estimates of a specific non-market good. Furthermore, the benefit estimates from the meta-analysis can be used to assign a value to a similar site if it is not possible to conduct an individual study on that site. This process in environmental valuation is called "benefit transfer."

A meta-analysis of the benefits of outdoor recreation in the United States between 1968 and 1988 found that 64.7% of 131 travel cost studies included term for substitutes, and the omission of term caused a 30% overestimation of the benefits to be obtained from the recreation (Walsh, Johnson & McKean, 1992). Shrestha and Loomis (2001) extended the research with regard to the United States outdoor recreation benefits and found that between 1967 and 1998 only 25.8% of 131 studies using both the travel cost and the contingent valuation methods, included substitutes in their model. A meta-analysis of the benfitts ecosystem services provided by lakes all over the world found that only 16% of 133 studies conducted between 1972 and 2012 utilizing all valuation methods – hedonic prising (HP), travel cost (TC), contingent valuation (CV), and choice experiments (CE) – included substitutes in their model (Reynaud & Lanzanova, 2017). A meta-analysis of the benefits of coastal recreation sites in Europe found that 63 of 177 observations from 38 valuation studies utilizing contingent valuation, travel cost, and choice experiments methods, also accounted substitute sites in their valuation (Ghermandi, 2014).

Based on these meta-analyses that concentrate on recreation, it is clear that only a small percentage of valuation studies include substitutes in their model. It is difficult to say how many studies focusing on water recreation have included substitutes in their models. Ghermandi (2014) is an exception, since it is the only meta-analysis focusing specifically on water-related recreation, but only with regard to European coastal areas. Other meta-analyses focusing on recreation have also included other types of recreational activities, such as hunting and winter sports. Similarly, since all the meta-analyses

included studies using various valuation methods (HP, TC, CV, and CE), it is difficult to distinguish which valuation methods have accounted for substitutes. The study by Walsh et al. (1992) represents an exception, because in their analysis, consisting of valuation studies employing the contingent valuation and travel cost methods, only the travel cost studies included term for substitutes.

2.3 Travel cost method and substitutes

The travel cost method stems from the simple observation that people are willing to pay to access recreational sites. This insight was first introduced by Hotelling (1947) in relation to valuing national parks. The monetary value of recreational services is elicited from the travel costs to the site, for example gasoline or bus tickets, and thus the travel cost method is based on revealed preferences. It is the most widely used environmental valuation method to estimate the monetary value of recreational activities. This thesis uses the travel cost method to value the recreational opportunities of the Baltic Sea.

Travel cost models can be roughly divided into single-site and multi-site applications. The individual travel cost model and the zonal travel cost model are single-site applications of the travel cost method, while the random utility model (RUM) is a multi-site application of the travel cost method. Single-site travel cost methods are used to obtain the total value of the site under valuation and RUM is used in situations where the research focuses on the value of environmental change at a site rather than the total value of the site.

The main difference between multi-site and single-site models is that in single-site applications the individual chooses the number of trips taken to the site under valuation and in RUM the individual chooses the site to visit among a set of possible sites instead of the number of trips to the site under valuation. The choice is influenced by the characteristics of the different sites. For the details of the random utility model see the example of Parsons (Parsons, 2003, pp. 296-302).

The distinct difference between the two types of single-site travel cost models is the predicted outcome of the models. In the single-site zonal travel cost model the area surrounding the site under valuation is divided into zones, usually in concentric circles surrounding the site. The model predicts the number of trips undertaken by the population of a particular zone within a certain time frame. The individual travel cost model predicts the number of trips taken by an individual within a certain time frame. The zonal travel cost method is suitable in a situation where the site under valuation is only visited once a year and the individual travel cost method is suitable in a situation where respondents undertake numerous visits. The appropriate travel cost method application depends on the valuation task at hand. In the case of the present research, the appropriate method is the

individual travel cost method, which is presented in the methodology section of this thesis. For details of the zonal single-site travel cost model, see for example Haab and McConnell (2005, pp. 181-182).

Rosenthal (1987) was one of the first to analyse how the welfare gain from a water recreation at a reservoir is impacted by substitute sites using the single-site travel cost model. His model included travel costs to substitute sites and revealed that omitting substitute sites caused significant overestimation of the welfare of the recreation site.

It is therefore suggested that the single-site travel cost model should include travel costs to substitute sites in the model (Freeman III, Herriges & Kling, 2014). The following are suggested ways to select substitute sites: 1) sites frequently visited by respondents, 2) sites that have similar characteristics to the study site, and 3) sites close to the study site (Parsons, 2003).

Recent single-site travel cost studies on the recreational value of different types of waterbodies have included substitute sites in various ways. A valuation of Australia's Gold Coast beaches included a substitute in the model as a dummy variable, indicating the existence of a substitute site for those respondents who had a favourite beach outside the study area (Zhang, Wang, Nunes & Ma, 2015). A study on the recreational value of three beaches in Spain took a similar approach, but the substitute variable was defined as a travel cost to an alternative beach site. The substitute was determined on the basis of respondents' choice of which beach they would visit if not visiting one of the study sites (Alves, Ballester, Rigall-I-Torrent, Ferreira & Benavente, 2017). Boyer, Melstrom and Sanders (2017) chose a substitute site for a reservoir in Oklahoma, being the most visited lake among respondents in a state area, and included travel costs to the substitute site in their model.

Common to these recent single-site studies is the inclusion of only one substitute site to the analysis similar to the site under valuation. However, two different practices were used to include substitute sites in the econometric models: 1) the travel costs to a user-identified site, and 2) the distance to a substitute site.

Most of the research used to assess the impact of substitutes on the welfare estimates of recreation has been conducted using the RUM. The reason for using the RUM is the "built-in" way of accounting for various substitutes for the site or the recreational activity under valuation. The RUM model requires that the analyst compiles a choice set of possible alternative sites or recreational activities for respondents to choose the option that maximises utility. There are two options for compiling a choice set: 1) include all possible alternatives that the researcher has considered, or 2) define a subset of substitute sites that are most likely relevant to the recreationist (Schaafsma, 2011, p. 69).

Based on economic theory, the choice sets used in the RUM estimation should include all relevant substitutes available to the individual (Hicks & Strand, 2000). In water recreation studies, the number of possible alternative sites increases as the distance to the site under valuation increases. Choice set formation is a difficult task for analysts, because only the respondent knows which substitutes are relevant to his/her decisions on recreation. A few studies have focused specifically on substitute sites or recreational activities that impact on welfare estimates, but all studies focusing on choice set formation have examined substitutes at some level.

The erroneous determination of a choice set in RUM modelling impacts the welfare estimate obtained from the analysis, because the welfare estimate is an explicit function of the choice set (Hicks & Strand, 2000). Omitting possible substitute sites or activities from the model leads to an underestimation of the welfare estimation if the study assesses the value of an environmental program that aims to improve the environmental quality (Jones & Lupi, 1999). By comparison, if the environmental quality decreases to the level that recreation is not possible at the study site, omitting substitute sites or activities leads to an overestimation of the welfare losses (Parsons, Plantinga & Boyle, 2000; Peters, Adamowicz & Boxall, 1995; Jones & Lupi, 2000). However, substitutes might not be the most important factor influencing welfare estimates, because an individual's welfare only changes if the site affected by environmental changes is in his/her choice set, thus it is possible that omitting substitutes only has a minor impact on welfare estimates (Hicks & Strand, 2000).

Analysts have used various ways to define choice sets and thus determine which substitute sites are relevant to respondents. Choice sets based on respondents' own perceptions are called endogenous choice sets. These types of sets are based on the idea that if an individual is not aware of a site, a quality change of the site does not affect the individuals' welfare. Hence the site is irrelevant to the individual and should be excluded from the analysis. The broadest definition of an endogenous choice set is that if an individual has any knowledge of a site, it is relevant to the analysis (Parson et al., 2000). Haab and Hicks (1997) have narrowed the concept of endogenous choice sets to an individual's knowledge of recreation sites where he/she can engage in his/her preferred recreational activity. In Peters et al. (1995), the endogenous choice set was restricted to visitation and intended visitation of recreation sites. They also considered the factors influencing awareness of possible substitute sites and found that awareness increased as expenditure related to the recreational activity and years engaged in the activity increased.

Choice sets formed exclusively by analysts are called exogenous choice sets. Parsons and Hauber (1998) included all possible lakes and rivers of a state area in their choice set and found that sites

added after a certain distance do not impact on the welfare measures of anglers, and therefore the sites further away from the site under valuation seemed to be irrelevant to respondents.

Common to all these studies aiming to define choice sets and relevant substitutes for the environmental amenity under valuation is that they only included very similar sites or recreational activities in the analysis. Also, the choice sets, and hence the possible substitutes, were defined by the analysts. This is even the case for the endogenous choice sets, where respondents could only choose from alternative sites outlined by the analysts. The limitation of this approach is that it overlooks the possibility that different types of areas or activities can be substitutes, for example, a forest area can be a substitute for a lake and boating can be a substitute for rock climbing.

The so-called freedom to roam laws or, as they are known in the Nordic countries, "everyman's right", leads to a unique problem related to RUM modelling. In Finland "everyman's right" guarantees open access to nature that is not dependent on ownership, for example, free movement in forest areas and swimming or angling in water systems are allowed (Tuunanen, Tarasti & Rautiainen, 2012). The open access situation in Finland and in all countries having similar laws leads to two separate issues: 1) the lack of distinct nature regions or sites, and 2) the number of substitute recreation sites is even greater (Ovaskainen et al., 2011). In these conditions, the formation of a choice set is difficult or even impossible.

2.4 Substitutes in stated preference studies

In stated preference studies, substitutes have been analysed in the context of distance decay. Distance decay is a well-known phenomenon were willingness to pay (WTP) for an environmental good under valuation decreases as distance to the site increases. One suggested explanation for distance decay is the existence of substitutes. The expectation is that as the distance to the environmental good under valuation increases, the number of possible substitute sites also increases, and hence distance decay and substitutes are interdependent. According to Rolfe and Windle (2012), environmental valuation has also revealed other explanations for distance decay; the further from the resource an individual lives, the lower his/her usage of it is and the less aware and responsible the individual feels for it. Stated preference analysis has mainly focused on how distance to substitutes impacts on the willingness to pay for an environmental quality change.

Pete and Loomis (1997) analysed WTP for three environmental programs aiming to improve environmental conditions. In two of the three cases in their study, WTP declined if respondents had substitutes in the proximity of their homes. De Valck, Broekx, Liekens, Aertsens and Vranken (2015) presented similar results in a study considering how the density of nature around respondents' homes

affected the value of nature sites under valuation. They had four different ways to include the distance to substitutes in their model.

Jørgensen et al. (2013) approached the impact of substitute sites on WTP for the improvement of the environmental condition of a Danish river by including travel time to the study site and two variables for substitutes in their model. The first variable was travel time to the nearest similar substitute site (stream or river) and the second variable was travel time to a less perfect substitute (coast). The major result of the study was that the proximity of substitutes decreased WTP for river restoration, but only for non-users. The suggested reason for this is that users did not consider the substitute sites chosen by the researcher as relevant substitutes for the site under valuation.

Schaafsma, Brouwer, Gilbert, Van Den Bergh and Wagtendonk (2013) tested the effect of substitutes on improving the environmental condition of three different waterbodies. They estimated two models, the first one including the distance to the site under valuation and the distance to one substitute site, the second including directional variables to analyse how the uneven spatial distribution of substitute sites affects WTP for the restoration. Their results indicated that omitting substitutes from the analysis can lead to an over- or underestimation of the WTP of both users and non-users. Their second model indicated that these results were also dependent on the direction from which the site is approached. Lizin, Brouwer, Liekens and Broeckx (2016) included substitute sites in their model, so that the distance to the first river under valuation was included in the model for the second river under valuation and vice versa, and found a slightly lower WTP for river restoration compared to the model that omitted substitutes.

The concept of a substitute is similar in stated preference literature to the concept of a substitute in revealed preference literature; in both substitutes chosen by analysts are used. The substitute sites are intended to be similar to the environmental good under valuation in terms of their environmental status and they services they offer (De Valck et al., 2015; Pete & Loomis, 1997; Lizin et al., 2016). Schaafsma et al. (2013) recognised that analyst-chosen substitute sites can be irrelevant to respondents. They used both respondents' answer-based substitutes as well researcher-chosen sites, but only the researcher-chosen sites were significant in their analysis. The reason for the insignificance of the respondents' identified substitutes is that these were small sites, not identified by the official databases from which the researcher-chosen sites were selected. Jørgensen et al. (2012) adopted a broader approach to the subject by considering that it is also possible that a site that is not similar to the study site can be a substitute. However, in their case, it was suggested that users did not consider the researcher-chosen substitutes relevant. De Valck et al. (2015) recognized the possibility that individuals' perceptions of what are important characteristics of nature are unique.

2.5 Who has substitutes: implications from the social psychology literature

Substitution as an aspect of recreational behavior, in other words, changing a preferred recreational site or activity for another, has been studied by social psychology. According to Gentner and Sutton (2008), both economics and social psychology have analysed the same phenomena based on similar assumptions about human behavior (that is, individuals make choices that maximize utility), but the modelling and concepts of the two disciplines are slightly different. The main difference is that social psychologists employ different contexts of substitution: activity substitution, site substitution or, for example in the context of fishing, substitution of the target species. Economists would consider all these as the same and the focus of economics is the impact of substitution on the welfare for the individual. By contrast, the monetary value of the welfare impact of substitutes is not part of social psychology research.

A study by Ditton and Sutton (2004) found that anglers' willingness to change to another recreational activity was influenced by demographics (age, education and gender) and their motivation to participate in fishing (only activity- specific motivations influenced them). Similar results were also found in a study of anglers substituting their preferred species for another species (Sutton & Ditton, 2005). Besides demographics and motivations to participate in certain recreational activities, place bonding via recreational specialization has also been found to influence anglers' willingness to substitute (Oh, Sutton & Sorice, 2013). Sutton and Oh (2015) found a strong link between commitment to fishing as a recreational activity and ability to change to another recreational activity. Han, Noh and Oh (2015) found similar results with anglers in a study of coastal tourism substitution behavior; motivations, place bonding and demographics influenced substitution behavior. Added to these, they also found a link between the condition of a preferred site and substitution. The driving forces behind substitution behavior are constraints individuals face while participating in recreational activities, for example lack of time, money, access and equipment. The constraints individuals face are at least influenced by demographics and motivations to participate in a certain recreational activity (Sutton, 2007).

Social psychology has found that demographics, motivation to participate in a certain recreational activity, place bonding, commitment to a preferred recreational activity and site condition influence individuals' willingness and ability to substitute. As we are also here interested of the question of who has a substitute site, we use these concepts from the social psychology literature to define possible relevant independent variables for our analysis of the question: "Who has substitute sites?".

3. Data

3.1 Survey implementation

The data used in this thesis were collected for the BalticApp project between November 2016 and February 2017. The BalticApp project is an EU project that ran from 2015 to 2018 and investigated supply and demand of the Baltic Sea ecosystem services in terms of the current projections of socio-economic development and pressure caused by climate change in the Baltic Sea. The main objective of the project was to identify strategies to ensure the supply of the ecosystem services for the future.

As part of the project, Latvians, Germans and Finns replied to the web-based survey about their recreational behavior at the Baltic Sea. Finnish and German respondents replied to the questions without additional assistance, but to ensure a representative sample of Latvians, personal interviews were conducted with them, because internet availability is not as universal in Latvia as it is in Finland and Germany. A total of 4,800 responses were received from the three countries, with approximately 2,000 individuals completing the survey in Finland, approximately 2,000 in Germany and approximately 760 in Latvia.

3.2 The questionnaire

The survey consisted of six sections. The first section contained information about the Baltic Sea and an introduction to the survey. The second section sought information about respondents' recreational behavior using map-based questions. In the third section respondents were asked detailed information about their last visit to their most frequently visited site at the Baltic Sea. The fourth section questioned respondents about their perceptions of the ecological condition of the Baltic Sea and their expected future visiting behavior. This section is related to contingent behavior research and only the questions about the perceived condition of the Baltic Sea are relevant to this study. In the fifth section respondents stated their preferred future condition of the Baltic Sea and replied to questions about the ecosystem services of the Baltic Sea. This section can be used for the choice experiment study and only the questions related to ecosystem services are important for this study. In the final sixth section respondents replied to questions regarding their backgrounds.

For this study, the second section of the questionnaire, on respondents' behavior visiting the Baltic Sea, was the most important. Respondents stated the number of visits they had made to the Baltic Sea: this question formed the basis for the travel cost analysis. Other information gathered in this section included the timing of visits and the recreational activities respondents had participated in at the Baltic Sea. The timing of visitation allowed for the separation of non-users and users. Respondents who had visited the Baltic Sea within the last three years were considered users and the rest were considered non-users.

The use of map-based questions in the second section distinguished the survey from earlier research. In the map-based questions, respondents navigated to a location on the map and pinpointed a place on the map. Map-based questions facilitated gathering information about the location of the respondents' place of residence and their most preferred recreation site. This spatial information was used to calculate various distances: from the respondents' homes to the Baltic Sea coast, and the road and Euclidean distance from the respondents' homes to recreation sites.

In the next section, respondents described the environmental condition of their most frequently visited site and the travel information to arrive at their most frequently visited site, including mode of transport and approximate travel costs. In the final section of the survey, demographic information was gathered. The complete survey included 41 questions divided into six sections. The anticipated time to complete the survey was 20 minutes.

In the survey users were also asked to determine possible substitute site or sites for their preferred recreation site at the Baltic Sea. The question on substitute sites was also a map-based question. Respondents pointed to a location on the map where they would find a similar recreational experience to their Baltic Sea recreation site if the ecologic condition of the Baltic Sea deteriorated to a level where recreation would not be possible. Respondents who did not choose a location on the map could state their reason for not having a substitute site by answering a multiple-choice question with three options: "I would stay at home," "I don't know where I would go" and "other reason." Due to the way the question was formulated, only spatial information about substitute sites was obtained from the survey. Examples of the survey questions are presented in Appendix A.

3.3 Data from geographical information systems (GIS)

The spatial information about substitute sites was used to describe the location of and land use surrounding the substitute sites. It also enabled us to calculate the various distances to the substitute sites. The purpose of this analysis was to obtain more accurate information about the substitute sites than merely the existence and location of the sites. The Euclidean distances from the substitute sites to the Baltic Sea coast as well as to the preferred recreation sites were calculated.

Various databases were used to describe the location of the substitute sites as accurately as possible. Land use types surrounding each substitute site were analysed within 500 m of the substitute sites. The land type data were obtained from the Copernicus database coordinated by the European Environment Agency (EEA) (EEA, 2019a). The Corine Land Cover (CLC) inventory provided by the Copernicus service includes 44 different land use classes and covers 38 countries, including Germany, Finland and Latvia.

The Marine Regions database, maintained by the Flanders Marine Institute in Belgium, provides geographic information about water systems, for example the standardized names of water systems and the geographic location of these features (Flanders Marine Institute, 2019). The database covers the entire globe and enabled to determine whether substitute sites are located at the ocean or inland. The European Union's Natura 2000 network of protected areas was assumed to offer good substitute sites for the Baltic Sea (European Commission, 2018). The Natura 2000 database is maintained by the EEA (2019b). The database contains extensive information about the areas that are part of the Natura 2000 network, including the spatial and ecological status of each area. The database was used to identify substitute sites that are part of the Natura 2000 network.

3.4 Descriptive information about the sample and substitute site

Respondents

A large portion of respondents did not complete the entire survey and the selective responses to the survey generated multiple subsamples of respondents. The formation of these subsamples is illustrated in Figure 1. The question on the timing of their last visit to the Baltic Sea divided respondents into users and non-users. Furthermore, the map-based questions divided users into two subsamples, "substitute question" and "no substitute question," because a large portion of users did not respond to these questions. Figure 1 also illustrates the number of respondents who have substitute sites and the number who do not. The descriptive statistics of these subsamples are presented in Table 1.

Sample structure 1

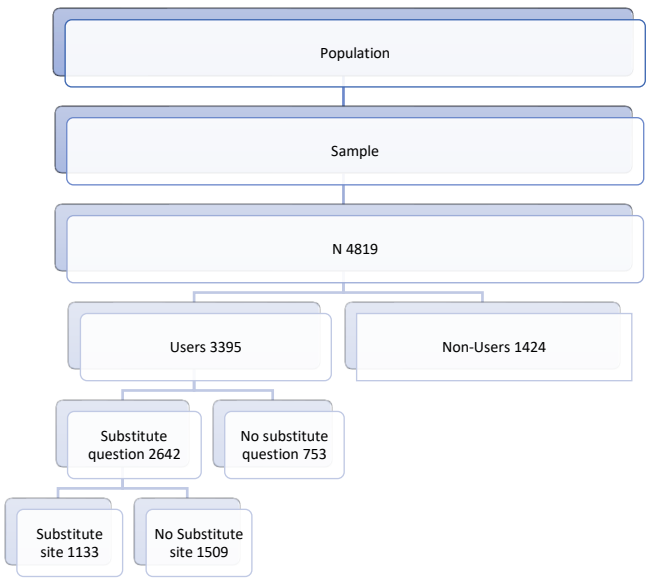


Figure 1

Table 1

Descriptive statistics of sample (means)

		All respondents	Users	Substitute question	No substitute question	Substitute site	No substitute site
Age		47,89	46,84	46,13	49,34	46,90	45,55
Income		1840,30	1897,29	1812,90	2189,40	2075,71	1605,16
Household size		2,24	2,32	2,32	2,31	2,31	2,32
Under 18-year-olds in a household		0,39	0,42	0,42	0,44	0,44	0,40
Gender	0=Male, 1=Female	0,51	0,51	0,50	0,53	0,50	0,51
Education	0=Lower 1=Higher	0,30	0,34	0,35	0,31	0,43	0,29
Occupation	0=Not working 1=Working	0,57	0,60	0,60	0,61	0,60	0,59
Finnish		0,43	0,46	0,48	0,41	0,52	0,44
Latvia		0,16	0,18	0,21	0,06	0,10	0,30
Germany		0,42	0,36	0,31	0,54	0,38	0,26
N		4819	3395	2642	753	1133	1509

Education dummy: The lower coded 0, means that respondent has compulsory education, vocational education or high school education and University/Polytechnic coded 1 means that respondent has university or polytechnic education.
Occupation dummy: Respondent is coded 0 if the occupational status is retired, student, unemployed or home-employed/homemaker respondent is coded 1 if the occupational status is employed full-time, part-time or self-employed.

A comparison of the descriptive statistics of the sample populations of Germany, Finland and Latvia indicates that the sample is fairly representative. The average population age for Latvia is 42.3, for Finland 42.5 and for Germany 44.3 (Statistics Finland, 2019; Central Statistical Bureau of Latvia, 2018; DesStat, 2019). The average age of our entire sample is slightly higher 47.89 years. The average population-level household sizes, according to the OECD, are 2.3 in Latvia, 2.1 in Finland and 2.0 in Germany (OECD, 2016). The average household size of the sample is 2.24, close to the population household sizes. The percentage of females of the entire population is 54.4% in Latvia, 50.7% in Finland and 50.8% in Germany (World Bank, 2017). This is similar to the sample, which has a slightly higher number of female respondents. The percentage of the population between the ages of 18 and 64 with a university/polytechnic degree in Latvia, Finland and Germany are 30.0%, 36.4% and 24.8%, respectively (Eurostat, 2019). Approximately 30% of the survey respondents have a university or polytechnic education, close to the population levels.

It is difficult compare occupational status at country level for two reasons: our data include individuals between the ages of 18 and 79 years and the employment levels of a country are measured for individuals between the ages of 15 and 64 years; and the definition employed also differs from our variable. Nevertheless, the employment rate in 2018 was 70.0% in Finland, 71.8% in Latvia and 75.3% in Germany (OECD, 2019). These are higher percentages than the sample percentage of 57%.

The same applies to the comparison of income levels in our data and the country-level incomes. The question on income in the survey employed eight income categories, and the mean income was calculated from the midpoints of these categories. Again, no real country-level measures for the accurate comparison of mean incomes exist, except for the household disposable income, which is 2371€ per month for Finland, 2345 € per month for Germany and 1330 € per month (OECD, 2016). These figures mostly reflect higher income levels than the sample income of 1840 € per month.

Turning to the comparison of the subsamples, the greatest variations among the subsamples are with regard to age, income level, education level and the respondents' home countries. Individuals who did not respond to the map-based questions are older than the "users" and those who answered the "substitute question." The "no substitute question" group has, on average, a much higher income level than the "users" or "substitute question" groups. Respondents who had a substitute site have a higher education level than the "users" group, but an even higher education level than the "no substitute question" group. There is also variation among respondents with regard to the home countries of the subsamples. Over half of the respondents in the subsample "no substitute question" are from Germany, compared to the other two subsamples, where approximately one third of respondents are from Germany. The portion of Latvians in the "no substitute question" subsample is also lower than in the other subsamples.

The second comparison between the "substitute site" group and the "no substitute site" group implies differences between those respondents that have a substitute site and those who do not. The most distinct differences between the "substitute site" and the "no substitute site" groups are that those respondents that have a substitute site have higher income and education levels.

Substitute sites

Of the 4,800 respondents, only 2,642 replied to the map-based questions, thus to the substitute question. Two factors limited the number of responses: 1) a portion of users had difficulty answering map-based questions; and 2) naturally non-users did not respond to the substitute question.

A total of 1,133 respondents identified at least one substitute site, and 181 of these respondents had more than one substitute site for the Baltic Sea. Altogether 370 substitute sites are located at the

Baltic Sea and the rest of the substitute sites are located inland. Information about the location is missing from 51 substitute sites. A total of 545 substitute sites were identified by respondents from Finland, 414 were identified by respondents from Germany and the rest were identified by respondents from Latvia. Altogether 224 substitute sites identified by respondents are Natura 2000 areas.

The mean distance from respondents' homes to substitute sites is 250.6 km, the median distance is 155.24 km, the minimum distance is 0.09 km and the maximum distance is 9,159.9 km. By comparison, the mean distance from respondents' recreation sites to substitute sites is 276.7 km, the median distance is 146.2 km, the minimum distance is 0.14 km and the maximum distance is 9,048.69 km.

There is no existing distance data for substitute sites. In a previous study, Lankia et al. (2013) calculated the mean distance to the typically visited swimming site and determined this to be 11 km. Possibly this reflects the lower boundary of the distance to the substitute site. The upper boundary of the distance to the substitute site is limitless, because a beach in Thailand could be a substitute site for the Baltic Sea. The median of the distance from home to the substitute site is relatively high, namely 155.24 km, but this could represent a trip to a summer cottage in Finland.

Land types surrounding the substitute site were divided into five classes: 1) artificial surface (towns, cities, etc.); 2) agricultural areas; 3) forest and semi-natural areas; 4) wetlands; and 5) water bodies (lakes, rivers, etc.). The distribution of major land types surrounding substitute sites was analysed for each country separately.

A total of 38.5% of Finnish respondents reported that their substitute sites are located near waterbodies. Artificial surfaces is the major land use, surrounding 20.4% of substitute sites, and 35.9% of substitute sites are located near forests. The distribution is somewhat different with regard to the substitute sites reported by German respondents: 41.9% of substitute sites are located near artificial surfaces. The major land type surrounding substitute sites is agricultural area, that is, in 25.6% of cases, with 14.8% of sites located in forests and 12.9% of sites located near waterbodies. Of the substitute sites identified by Latvians, 46.3% are located near waterbodies, 26.9% near artificial surfaces and 17.9% in forest areas.

The high percentage of artificial surfaces and forest areas being reported by the respondents of all countries as the main land use types surrounding substitute sites implies that water is not necessarily needed for the recreational activities at substitute sites. This is particularly the case with regard to German respondents, who reported that relatively few substitute sites are located at waterbodies, compared to the respondents from Finland and Latvia.

1. Statistical/econometric models

4.1 Logistic regression

The statistical method selected to analyse the question "Who has a substitute site or sites?" was logistic regression. The dependent variable of the analysis was dichotomous: either respondents had a substitute site or they did not. Logistic regression was applied to the group of respondents who completed the entire survey and replied to the map-based questions, as well as to the substitute question.

The estimation and analysis of the logistic model followed the steps introduced by Hosmer, Lemeshow and Sturdivant (2013, pp. 90-93). The first step was to devise a strategy for choosing independent variables for the analysis. There were two possible strategies: 1) to include all relevant variables even if the variables had not been statistically significant in the preliminary testing; and 2) to include only the variables that are supported by the theory and were statistically significant in the preliminary testing. The latter method is called purposeful selection and was used for this analysis.

Preliminary testing was performed with regard to each independent variable. The tests indicated whether there was a statistically significant relationship between each of the separate independent variables and the dependent variable. Pearson's chi-squared test (chi-squared test for independence) compares the observed frequencies of the data to the expected frequencies produced by the test. The expected frequencies are defined so that no relationship exists between the variables. The test is suitable for the categorical and dummy variables. One-way analysis of variance (ANOVA) was used for the continuous variables. An ANOVA indicates whether there is a difference between the means of the dependent variable groups.

Variables of the logistic regression

Before the final model was estimated, categorical variables (travel mode, education, age and site condition variables) were modified to dummy variables. This made interpretation of the final model easier. The dummy variables were based on Pearson's chi-squared test of independence. The turning point of the ratio between the expected values produced by the test and the observed values of data from higher to lower or vice versa was used as a line to assign the observations to a group, 1 or 0.

The independent variables for the logistic regression were selected by comparing the social psychology literature and the survey questions. These variables included demographics, the conditions of respondents' most frequently visited site and whether respondents had participated in a certain recreational activity at the Baltic Sea. Three variables could indicate place attachment, namely hours stayed at the most frequently visited site, visitation times at the most frequently visited

site and the importance of cultural and historic sites. The importance of recreation indicated a commitment to recreation. The distance variables were also assumed to be good indicators, because as the distance to the site under valuation increases, the number of substitute sites increases. A total of 41 different variables were identified, but based on the preliminary tests, only 20 variables were significant and included in the model. The descriptive statistics and expected signs of all the independent variables are listed in Appendix A and the significant variables are presented in Table 2.

Table 2 variables for the logistic regression

Dependent variable		Mean	Std. Dev.	Min	Max	n
Substitution	0=No substitute site, 1= Substitute site	0,43	0,50	0	1	2642
Independent variables						
Timing of visitation	0=Over 12 months ago, 1=Inside 12 months ago	0,76	0,43	0	1	2642
Number of activities	0=One, 1=Two or more	0,74	0,44	0	1	2544
Time	0,05-700 h	39,04	70,46	0	700	2621
Importance of recreation	0=0-25, 1=26-100	0,54	0,50	0	1	2641
Walking	0=No walking 1=Walking	0,68	0,47	0	1	2544
Jogging	0=No jogging, 1=Jogging	0,08	0,27	0	1	2544
Swimming	0=No swimming, 1=Swimming	0,32	0,47	0	1	2544
Nature watching	0=No nature watching, 1=Nature watching	0,31	0,46	0	1	2544
Diving	0=No diving, 1=Diving	0,01	0,07	0	1	2544
Boating	0=No boating, 1=Boating	0,07	0,25	0	1	2544
Sunbathing	0=No sunbathing, 1=Sunbathing	0,24	0,43	0	1	2544
Picnicking	0=No picnicking, 1=Picnicking	0,20	0,40	0	1	2544
Sauna	0=No sauna, 1=Sauna	0,04	0,20	0	1	2544
Travel mode	0=Public transport, 1=Car	0,79	0,41	0	1	2189
Number of species	0=Low and rather low, 1=Rather high and high	0,61	0,49	0	1	2096
Number of facilities	0=None or some, 1=Many	0,45	0,50	0	1	2497
Age	0=Between 18-29 and over 65, 1=Between 30-64	0,68	0,47	0	1	2642
Education	0=Lower education, 1=University/polytechnic	0,35	0,48	0	1	2641
Finnish	0=Not Finnish, 1=Finnish	0,48	0,50	0	1	2642
Latvian	0=Not Latvian, 1=Latvian	0,21	0,41	0	1	2642
German	0=Not German, 1=German	0,31	0,46	0	1	2642
Income (continuous)	50-5000€	1812,90	1237,22	50	5000	2281
Distance from home to recreation site	0-1736,25 km	154,62	185,08	0	1736,25	2609

The replies to the survey follow anticipated distributions among the respondents who belong to the sub-sample "substitute question." These respondents answered the entire survey, including map-based questions. A large share of our sample has visited the Baltic Sea within 12 months and 70% of the respondents participate in more than one recreational activity at the Baltic Sea, highlighting the fact that the Baltic Sea is an important recreation site with many recreational opportunities for Finns, Latvians and Germans. Activities also follow the anticipated distribution, with most of the respondents participating in walking, sunbathing, picnicking and swimming. These are low-cost, easily accessible activities compared to diving, boating and taking a sauna. The site condition variables do not represent absolute numbers but rather the perceptions of individuals of the most preferred sites. Based on these variables, respondents' most frequently visited sites are in a good condition. About half of the respondents reported that the number of species at the site are high and there are many facilities at their most frequently visited site. These percentages could also be lower due to the fact that the ecological condition of the Baltic Sea is relatively poor (Helcom, 2018). The mean Euclidean distance from respondents' homes to the recreation sites is approximately 154 km, indicating that a high number of respondents do not live along the coast but are still a day trip away from the Baltic Sea.

The respondents indicated the travel mode used to arrive at their most frequently visited site in response to a multiple-choice question listing seven possible modes: 1=Walk, 2=Bike, 3=Car, 4=Public Transport, 5=Private Boat, 6=Ferry, 7=Other. Based on the literature, it was assumed that individuals arriving with cars at the most frequently visited site would have access to more sites than individuals limited by public transportation. The travel mode dummy was coded 1 if a respondent arrived by car and 0 if a respondent used public transport.

The original site condition variables were divided into categories. Respondents rated the condition of their most frequently visited site on a 5-point rating scale. The number of perceived species was measured on a 4-point rating scale, ranging from low to high, and the number of perceived facilities at the most frequently visited site was measured on a 3-point rating scale, ranging from none to many. The ratio between the expected and the observed values determine the formulation of the site condition dummy variables. With regard to the number of perceived species variable, responses of rather high and high were assigned to group 1, and responses of rather low and low were assigned to group 0. With regard to the number of facilities variable, responses of none and some were coded as 0 and high was coded as 1.

Respondents stated their level of education in response to a multiple-choice question, with the four categories being compulsory education, vocational education, high school and university/polytechnic education. Similar to the earlier question, the ratio between the expected and the observed values

determined the formulation of the education dummy. Lower than high school education was coded as 0 and university/polytechnic education was coded as 1.

Age categories were based on the responses to an open-ended question in the survey and the four categories were formulated according to the classification of Statistics Finland (2019). The ratio between the expected and the observed numbers was used to formulate the final variable, were respondents younger than 18-29 and respondents older than 65 were coded as 0 and respondents between 29 and 65 were coded as 1.

In addition to the travel mode, site condition, level of education and age, the other significant variables are depicted in detail below. Respondents indicated the recreational activity or activities in which they participated at the most frequently visited site at the Baltic Sea. Variables were labelled by the distinct name of the recreational activity were coded as 1 if respondents participated in the activity and as 0 if they did not. The number of activities dummy was formulated by summing up the responses to the single recreational activity questions. With regard to the number of activities dummy, participating in more than one activity at the Baltic Sea was coded as 1 and participating in only one activity was coded as 0.

The timing of visits to the Baltic Sea as well as the length of visits were captured in the survey. In the analysis, these variables are named “timing of visitation” and “time.” Respondents who had visited the Baltic Sea within the last 12 months were coded as 1 and respondents who visited more than 12 months ago were coded as 0. The hours spent at the most frequently visited site constitute a continuous variable ranging from 0,05 to 700 hours.

Respondents indicated the importance of various ecosystem services contributing to their personal value measurement of the Baltic Sea. In the constant sum question, respondents distributed 100 points to eight ecosystem services available at the Baltic Sea, for example opportunities for recreational activities, habitats for many animals and inspiration for artistic work. The question can be found in Appendix A. Points distributed to the opportunities for recreational activities formed the basis for the importance of recreation. In the dummy variable, low points ranging from 0 to 25 were coded as 0 and high points between 26 and 100 were coded as 1. Low points indicate that recreation is not important to the respondents and high points indicate the opposite.

Respondents indicated their level of income in terms of one of eight categories where 1=less than 200 euros, 2=201-500 euros, 3=501-900 euros, 4=901-1600 euros, 5=1601-2500 euros, 6=2501-3600 euros, 7=3601-5000 euros, 8=over 5000 euros . The continuous income used in the analysis was based

on these categories, midpoints of these categories were user to compile the continuous income variable.

According to the preliminary testing, the only significant distance variable was the analyst-calculated distance from respondents' homes to recreation sites, ranging from 0 km to 1,100 km.

Estimation of the logistic regression

The logistic regression model can be represented as follows (Menard & Menard, 2010, p. 19; Hosmer et al., 2013, p.35):

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + .. + \beta_k\chi_k$$

where P=probability of having a substitute site, (p/1-p) = odds of having a substitute site, β_0=constant, β_k=parameter estimate and x_k=independent variable.

Before calculating the estimation, it was ensured that the assumptions of the logistic regression were not violated (Menard & Menard, 2010, pp. 125-126). One of these assumptions is that correlation of the independent variables, known as collinearity, is not allowed. Collinearity leads to biased coefficients (the estimated coefficients are systemically too high or too low), inefficient estimates (the coefficients have large standard errors) and invalid statistical inference (the statistical significance of the regression coefficients is inaccurate). To avoid collinearity issues, correlations of independent variables were tested with Pearson's correlation coefficient (bivariate correlation) (Hosmer et al., 2013, p. 182). There were two significant correlations between independent variables that could cause collinearity issues. "Number of activities" was correlated with variables indicating participation in certain recreational activities at the Baltic Sea. Depending on the activity, the correlation varied between 0.017 and 0.339. The second correlation issue was between "timing of visitation" and "distance from respondents' homes to recreation site;" this correlation was 0,284.

Reliable estimates for the parameters were obtained by using maximum likelihood estimation. The likelihood function is constructed for an estimation that contains all relevant information from the data. The maximization of the likelihood function results in the highest probability of obtaining the data from the independent variables. See Agresti (2013, pp. 134-135) for the likelihood function for logistic regression.

To avoid collinearity issues, six different models were estimated. Stepwise estimation was performed for all of the models and, based on the univariable Wald test, all insignificant variables were dropped one at a time (the most insignificant first). Based on the Wald test, the independent variable "distance from respondents' homes to recreation site" was insignificant in all of the models; because of this,

most of the estimated models shrunk to consist of identical independent variables. After the estimation, only two distinct models remained: 1) the model including the variable "number of activities"; and 2) the model including variables indicating participation in certain recreational activities. The other independent variables in the models were the same. The first model was selected as the final model, because of the more parsimonious nature of this model.

Goodness of fit of the logistic regression

A number of tests, known as goodness of fit tests, can be used to assess the logistic regression model's predictive capabilities. These are the likelihood ratio test, the Wald test, the pseudo R^2 values and the Hosmer-Lemeshow test. The model's classification table is also considered to be a good evaluation method.

The likelihood ratio test compares the likelihood of the fitted model to a null model. The Wald, or score, test is similar to the likelihood ratio test and can be used for identical comparison (Hosmer, 2013, pp. 12-15). Significant test results for the likelihood ratio test or the Wald test indicate that the final model offers a statistically significantly better explanation of the outcome than the baseline model (no independent variables, only the intercept).

There are many different coefficient of determination values for logistic regression models, sometimes referred to as "pseudo R^2," that compare dependant variable outcomes with the continuous predicted values that the model produces. Menard (2000) compared five pseudo R^2 values for the logistic regression. He stated that the major issue with all pseudo R^2 values for logistic regression is the lack of intuitive interpretation. Added to this, the usually reported pseudo R^2 value of the Cox & Snell R^2 cannot reach a value of 1. The issue is fixed by Nagelkerke R^2, an adjusted version of Cox & Snell R^2. Even with the issue of interpretation, the closer to 1 the pseudo R^2 values are, the better the model fits the data.

According to Menard (2013, pp. 57), the Hosmer-Lemeshow test indicates how the predictions of the model fit the dependent variable group memberships. The test compares the observed and predicted values of the model. An insignificant test result indicates that the model fits the data. However, the Hosmer-Lemeshow test is sensitive to the sample size. With small samples (n<400), the test fails to pick up small misspecifications of the model, and with large samples (n in thousands), even a small difference between the predicted and observed values can cause the test being statistically significant and indicating that the model does not fit the data well (Hosmer et al., 2013, pp. 158-168).

The model classification table is a similar assessment of the data to the Hosmer-Lemeshow test, but no econometric tests are associated with the classification table. The overall predictive capability just indicates how the model predicts the dependent variables outcome.

Interpretation of the logistic regression

The logistic regression is interpreted in terms of the odds ratio (OR). The OR approximates how likely or unlikely some outcome is in terms of the odds. The OR for a single variable in logistic regression is calculated from the confidents of the model $OR = e^{\beta_i}$. The interpretation is straightforward if the independent variable is a dummy variable: If the coefficient of the independent variable is larger than zero, the group coded 1 in the independent variable has greater odds of having the dependent variable outcome coded 1 than the group coded 0 in the independent variable. By comparison, if the coefficient is less than 0 belonging to the group coded 1 in the independent variable, this decreases the odds of having the dependent variable outcome (Hosmer et al., 2013, p. 50). In other words, an independent variable having a negative sign indicates that the group coded 1 in the variable has smaller odds of having a substitute site than the group coded 0. A positive sign indicates that the group coded 1 in the variable has greater odds of having a substitute site than the group 0. The more negative the coefficient parameter is, the smaller the odds are and, by comparison, the more positive the coefficient parameter is, the greater the odds are. The results of the logistic regression are presented in Table 5.

4.2 Single-site travel cost model

Model and its issues

The individual travel cost model was used to assess the recreational value of the Baltic Sea, as well as the impact of substitutes on welfare recreation at the Baltic Sea. The individual travel cost model has a firm basis in consumer theory (Phaneuf & Smith, 2005, pp. 683-685). The final product of the single-site travel cost model is downward sloping in terms of the demand function of recreation, where the quantity demanded is the number of trips taken to the site under valuation and the price is the travel costs to the site. As stated earlier, the individual single-site travel cost predicts the number of trips (r) taken to the site under valuation. The formulation presented here follows Parsons (2003) and Haab and McConnell (2005).

Most simple single-site models include only travel costs to the recreation site tc_r and the income *y* of the individual. Income is included in the model because it is one of the main shifters of demand and the positive effect of income on the demand for goods or services has a strong basis in economic theory, indicating that the good is a normal good.

$$r = f(tc_r, y)$$

The model was further expanded because the demand function for recreation also depends on other factors; by adding the price of substitutes tc_s, model includes another important shifter of demand. Besides travel costs, income and substitutes, other good predictors are demographics *z* and case sensitive factors *q*, for example, the site condition.

$$r = f(tc_r, tc_s, y, z, q)$$

Besides substitutes, there are several unresolved issues in the presented travel cost model, namely how to include the opportunity cost of time, multiple-destination trips and length of visit in the analysis.

One of the ongoing debates in travel cost analysis is how to include the opportunity cost of time (travel time to site) in the travel cost variable. Depending on individual perceptions of the actual trip, omitting the travel time from the travel cost parameter causes a downward or upward bias in the welfare estimate (Ward & Beal, 2000, pp. 34-38). The individual has to give up part of their income to travel to a recreation site: this trade-off between work and recreation time forms the theoretical basis for the travel time issue (Parsons, 2003). The well-established way in recreation demand is to use one third of an individual's full wage as a good estimate for the opportunity cost of time and to use analyst-calculated travel costs. Ovaskainen, Neuvonen and Pouta (2011) found that perceptions of the actual trip varies between respondents with regard to cost and welfare. They also compared respondents' perceived travel costs to the conventional wage-based travel cost and found only a slight difference in the welfare estimate. Similarly, Hagerty and Moeltner (2005) found that the difference in welfare estimates was not significant between respondent-perceived and analyst-calculated travel costs, although the difference in travel costs existed. In recent water-based recreation studies, the conventional wage-based approach was used (Hanauer & Reid, 2017; Zhang et al., 2015; Alves et al., 2017), but some studies have excluded time costs completely from their travel cost parameters (Akron et al., 2017; Hynes et al., 2017). Freeman lll et al. (2014, pp. 296-300) reviewed the issue of using one third of wages as an estimate for the opportunity cost of time and also presented alternative practices to include the opportunity cost of time in the analysis.

The second issue related to the opportunity cost of time is the treatment of on-site time. The exclusion of on-site time from an analysis causes the omission of benefits acquired staying at the site. On-site time is usually correlated with travel costs, possibly causing collinearity issues in the model. Correlation stems from a natural assumption: individuals travelling from farther away to the site also tend to stay longer at site. Some recent studies on water recreation do not account for on-site time at all (Zhang et al., 2014; Alves et al., 2017). Hynes et al. (2017) had a separate variable in their model

indicating number of days stayed at the site. The Latinopoulos (2014) model had a dummy variable indicating the difference between single-day visits and longer visits.

The basic assumption of the travel cost model is that recreation is the only purpose of trips taken to recreation sites. However, trips to recreation sites can have multiple purposes, for example also visiting friends or family. The welfare estimate is possibly biased if the only-purpose assumption is violated. The treatment of multi-purpose trips can result in bias. Martínez-Espiñeira and Amoako-Tuffour (2009) have stated that approaches in the travel cost literature vary: the complete exclusion of multi-purpose visitors from a sample causes an underestimation of the welfare of recreation, while ignoring the issue by treating all trips as only-purpose trips causes an overestimation of the welfare of recreation. They therefore provided the currently used correction of the multiple-purpose issue by weighting the travel cost parameter based on respondents' trip intentions. In recent travel cost studies, the weighted approach is used (Preez & Hosking, 2011; Preez, Dicken & Hosking, 2012). No correction of this issue is needed if a large majority of respondents only visited the site under environmental valuation (Mangan et al., 2014). Some recent studies have completely excluded multi-purpose trips from their analyses (Zhang et al., 2014; Hynes et al., 2017). Parsons and Wilson (1997) introduced a simple dummy variable to the model, because the calculation of accurate costs for multi-purpose trips was too complicated. They found that models including a correction dummy produced more conservative welfare estimates compared to models completely omitting multi-purpose trips. In recent water recreation studies, Akron et al. (2017) analysed multi-purpose trips with a dummy variable.

Poisson model

Count data models are suitable for the estimation of recreation demand. The non-negative integer character of dependent variable trips to the Baltic Sea can be captured by count data models. Poisson distribution forms the basis for modelling counts, but the assumption of an equal mean and variance called equidispersion is the drawback of Poisson distribution, as equidispersion is rarely the case with real data. Even though it is not used regularly in recreation demand analysis, Poisson distribution is presented here for two reasons: 1) it forms the basis for modelling count data; and 2) the sample selection model for count data presented later in this thesis is based on Poisson distribution. The unconditional Poisson probability density function indicates the probability of an individual taking a trip to the Baltic Sea, where μ_i is the predicted mean count of the number of trips taken, y_i is the number of trips taken and $y_i!$ is the factorial of the number of trips taken (Hilbe, 2011, p. 80).

$$\Pr(y_i;\mu_i) = \frac{\mu_i^{y_i}\exp(-\mu_i)}{y_i!} = \frac{\mu_i^{y_i}e^{-\mu_i}}{y_i!}$$

Negative binomial model

The negative binomial model is chosen as the count data model over the Poisson model for the single-site travel cost model in this thesis. Some form of negative binomial model is used for modelling in many travel cost studies (Akron et al., 2017; Boyer et al., 2017). The negative binomial model is particularly suitable for analysing recreation demand for several reasons. First, similar to the Poisson model, only positive integers are allowed in the dependent variable. Trips taken to the recreation site are exactly that. Second, the negative binomial model has an advantage over the Poisson model in that it allows overdispersion in the data, meaning that the conditional mean and the conditional variance of the dependent variable do not need to be equal, precisely the variance is greater than the mean (Freeman lll et al., 2014, p. 279; Haab & McConnell, 2005, p. 169).

Following Winkelmann (2008), Hellerstein (1991) and Haab and McConnell (2005), the probability density function of the negative binomial model, thus the probability of an individual taking trip x_i in a certain timeframe, is calculated as follows:

$$\Pr(x_i) = \frac{\Gamma\left(x_i + \frac{1}{\alpha}\right)}{\Gamma(x_i + 1)\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \lambda_i}\right)^{\frac{1}{\alpha}} \left(\frac{\lambda}{\frac{1}{\alpha} + \lambda_i}\right)^{x_i}, x_i = 0, 1, 2, \dots$$

where Γ is the gamma function. The parameter alpha α in the negative binomial distribution is called the overdispersion parameter. As stated earlier, trips taken to the recreation site are positive integers. To ensure this condition, the exponential function is chosen for the expected value λ_i. The expected value is the expected number of trips taken by an individual. The mean and the variance in the exponential form are $E(x_i) = \lambda_i = e^{(z_i\beta)}$ and $Var(x_i) = \lambda_i(1 + \alpha\,\lambda_i)$. More precisely, the expected value, following Edwards, Parsons and Myers (2011), is calculated as follows:

$$E(x_i) = \lambda_i = e^{(z_i\beta)} = e^{(\beta_{tc}tc_i + \beta_{tcs}tcs_i + \beta_y y_i + \beta_z z_i)}$$

This type of parametrization of the negative binomial model is particularly useful, since it follows that if parameter alpha α approaches infinity, the probability density function of the negative binomial distribution turns to the probability density function of the Poisson distribution. The Poisson distribution is nested in the negative binomial distribution. This link between the distributions is used when the negative binomial model is evaluated. The simple significant likelihood ratio test $\alpha = 0$ indicates that the distribution of trips is overdispersed and negative binomial distribution is appropriate for the model.

Variables of the travel cost model

The dependent variable of the travel cost model is the number of trips to the Baltic Sea. In the survey, respondents who had visited the Baltic Sea within three years were considered users and respondents who had visited over three years ago were considered non-users. Possible outliers were removed from the variable by limiting the number of trips to the Baltic Sea in one year to 365. Trips taken more than a year ago but less than three years ago were divided by three, to be equal to the number of trips taken within the last 12 months. To satisfy the requirement of only including positive integers in the dependent variable, trips were rounded off to the closest integer. After the rounding off, 265 zero observations remained in the dependant variable, because 265 respondents had visited the Baltic Sea only once within the last three years. These observations were removed from the dependant variable. The final result is that the variable ranges between 1 and 365 trips to the Baltic Sea during one year.

Travel costs, income and substitutes formed the basis of the independent variables. The other independent variables were demographics, site condition, the importance of recreation as indicated by respondents, the number of recreational activities in which respondents participated and the travel mode used to arrive at the most frequently visited site, which were all selected as good predictors of the number of trips an individual had taken to the Baltic Sea. The number of facilities, the number of activities, the importance of recreation, time, income, Finnish and Latvian were all variables already used in the logistic regression. Details of the variables selected for the analysis and the treatment of the opportunity cost of time and multi-purpose trip issues are presented below. Descriptive statistics of the independent and dependent variables are listed in Table 3.

Table 3 variables for the travel cost model

Dependent variable		Mean	Std. Dev.	Min	Max	n
Trips to the Baltic Sea		13,08	43,65	1	365	3125
Independent variables						
TC	0-1000€	94,98	154,56	0	1000	3011
TC*SUBS	0-1000€	33,52	100,62	0	1000	2558
Substitute	0=No Substitute site, 1= Substitute site	0,43	0,50	0	1	2642
Income	50-5000€	1897,29	1227,46	50	5000	2940
Age	18-79	46,84	15,31	18	79	3394
Gender	0=Male, 1=Female	0,51	0,50	0	1	3394
Occupation	0=Not working, 1=Working	60,1	0,49	0	1	3393
Education	0=Lower, 1=Higher	0,56	0,50	0	1	3394
Household size	1-8	2,32	1,14	1	8	3383
Finnish	0=Not Finnish, 1=Finnish	0,46	0,50	0	1	3395
Latvian	0=Not Latvian, 1=Latvian	0,18	0,38	0	1	3395
Water clarity	0=Turbid, 1=Clear	0,59	0,49	0	1	3061

Number of facilities	0=None/some, 1= Many	0,45	0,50	0	1	3180
Number of activities	0=One, 1=Two or more	0,74	0,44	0	1	3280
Importance of recreation	0=0-25, 1=26-100	0,52	0,50	0	1	3393
Travel mode	0=Other, 1=Walk/bike	0,10	0,30	0	1	3393
Multi-purpose trips	0=Recreation not the only purpose, 1=Recreation only purpose	0,42	0,49	0	1	3392
Time	0.05-700h	41,98	75,03	0,05	700	3358

The travel cost analysis included all users of the Baltic Sea. The distribution of the new variables introduced to the analysis are reasonably similar to the logistic regression analysis. The average number of trips taken to the Baltic Sea is ~13. This is not surprising, as the Baltic Sea is an important recreation site for Finns, Germans and Latvians. The average stated travel cost is approximately 94€, because of the long average distance to the most preferred recreation site: at ~154 km this is reasonable. Since the coastal areas of Latvia, Germany and Finland are densely populated, it is not surprising that more than half of the respondents had purposes other than recreation for their trips to the Baltic Sea. It is likely that respondents have relatives or friends living at the coast of the Baltic Sea. The average time spent at the recreation site is more than one day, but the maximum range of the time variable is ~29 days, meaning that a great number of trips are less than one day in duration. The numbers of respondents valuing recreation as an important ecosystem service of the Baltic Sea and the respondents not valuing recreation as an important ecosystem service are almost equal. This is not surprising, since the Baltic Sea provides a number of other important ecosystem services, for example, it is an environment for learning and gaining new information and an important habitat for plants and animals.

The construction of variables for the travel cost model began with the travel cost variable. This thesis excluded the opportunity cost of time and used respondent-stated travel costs in the travel cost variable (TC). The survey respondents reported one-way travel costs that included all other expenses except meals and accommodation. Outliers were excluded by limiting the responses of one-way travel costs to 500€. Round-trip travel costs were calculated by multiplying the one-way travel costs by two. The natural expectation is that the number of trips will decrease as the travel costs increase.

The model included on-site time as a continuous variable based on respondents' reported hours spent at the most visited site. The time variable is limited to 700 h to exclude outliers.

Substitute sites are included in the analysis as a dummy variable as well as substitute interaction term. The substitute dummy variable is coded 1 if a respondent has a substitute site and 0 if not. The

substitute dummy captures the impact of substitute sites on the number of trips taken. The expectation is that the sign of the variable is negative, if a respondent has a substitute site fewer trips are taken. The interaction term is labelled TC*SUBS. The interaction variable is constructed by multiplying the TC variable with substitute dummy variable, indicating whether a respondent has a substitute site. These types of interaction terms are used in travel cost models (Lankia, Neuvonen & Pouta, 2017; Huhtala & Lankia, 2012; Loomis, Gonzales-Caban & Englin, 2001; Bhat, 2003; Blaine et al., 2014). The interaction term TC*SUBS affects the slope of the estimated recreation demand curve. The change in the steepness of the slope indicates the effect of substitute sites on the welfare estimate. According to economic theory, price elasticity is higher for goods or services that have substitutes, thus the negative sign of the interaction. It is expected that the recreation demand curve will be steeper for individuals with substitute sites compared to those without, leading to lower welfare estimates.

The survey measured respondents' income in terms of eight categories. Categories are 1=less than 200€, 2=201-500€, 3=501-900€, 4=901-1600€, 5=1601-2500€, 6=2501-3600€, 7=3601-5000€, 8=over 5000€. The continuous income variable used in this analysis was based on the category midpoints. The expectation is that as income increases, the number of trips taken also increases (Parsons, 2003). By comparison, some variation was seen in water recreation analysis. One study found a negative impact (Huhtala & Lankia, 2012) and income was an insignificant predictor (Hynes et al., 2017; Hynes et al., 2015).

The survey included a multiple-choice question where respondents indicated whether they had multiple purposes for their trips to the Baltic Sea. Instead of using the weighted approach, this thesis used a simple dummy variable to indicate whether a visit to the Baltic Sea was the only purpose of a trip. Respondents were coded 1 if recreation was the only purpose of the trip and 0 if not. The expectation is that if recreation is the only purpose of a trip to the Baltic Sea, the total number of trips will decrease.

Generally, demographics have been considered good predictors in the travel cost model. The impact of demographics on the number of trips taken is highly case sensitive and depends on the valuation task at hand (Ward & Beal, 2000, pp. 71-73). For example, Alves et al. (2017) assessed the recreational value of three beaches in Spain and found that age and education can either increase or decrease the number of trips taken, depending on the visited beach. Demographics are included in this analysis in various forms, for example, continuous variables for age and household size. In the education dummy variable, education lower than vocational education was coded 1 and higher than high school education was coded 2. In the occupation dummy, employed were coded 1 and unemployed

individuals were coded 0. Retired persons, students and unemployed individuals were regarded as unemployed and full-time employed, part-time employed and self-employed individuals were regarded as employed. Dummy variables also indicated respondents' gender and home country.

Respondents indicated the perceived condition of their most frequently visited site on a rating scale. To avoid correlation, only two out of six measures were included in the analysis: current water clarity and number of facilities. Current water clarity indicates how deep respondents could see under the surface of the water. The number of facilities dummy is the same variable described in the logistic regression analysis. The current water clarity dummy coded respondents choosing "turbid" and "somewhat turbid" in the group 0 and those choosing "somewhat clear" and "clear" in the group 1. The expectation is that if respondents perceived having good water clarity and many facilities, they would take more trips to the Baltic Sea.

Besides the obvious predictors of travel costs, substitute sites, income, demographics and site condition, single-site travel cost studies include other variables that affect the number of trips. The selection of these variables depends on the data and the environmental amenity under valuation. This thesis includes variables indicating the travel mode to arrive at the most frequently visited site, the importance of recreation for the respondent and the number of recreational activities the respondent participates in at the Baltic Sea. The number of recreational activities variable and the importance of recreation have already been described in the logistic regression analysis. The expectation is that if a respondent participates in more than one recreational activity at the Baltic Sea or considers recreation important, the respondent will take more trips.

The multiple-choice question on the travel mode had seven options: car, public transport, private boat, ferry, walking, biking or other transport mode. The time cost of the trip was omitted from the analysis because of the assumption that travel costs are zero for bikers and walkers. In the travel cost dummy, bikers and walkers were coded 1 and all other travel modes were coded 0. The purpose of the dummy is to capture the different travel costs for bikers and walkers. Lankia et al. (2017) created a similar measurement, but they had separate travel cost variable for bikers and walkers.

Estimation of the travel cost model

Similar to logistic regression, consistent estimates for the parameters of the negative binomial model are achieved by using a maximum likelihood estimation (Hilbe, 2011, p. 190). A stepwise estimation was performed to arrive at the final model. Similar to logistic regression, insignificant variables were dropped, based on the Wald test.

To capture the impact of substitute sites on the number of trips taken, as well as on the welfare estimates, five models were estimated. The first model included only travel costs and income, two important shifters of demand. The second model introduced all other independent variables to the estimation except substitute variables. The third model included the substitute sites dummy variable and other independent variables, but excluded the travel cost substitute interaction variable from the estimation. Model four included the substitute travel cost interaction term and all other variables except the substitute dummy variables. Model five included all variables of the estimation. A stepwise estimation was performed to arrive at the final models. Insignificant variables were dropped, based on the Wald test. All five distinct models are presented in Table 6.

Interpretation and assessment of negative binomial model

The interpretation of the negative binomial model's coefficients depends on the type of the variable. The model can be interpreted using incident rate ratios (IRR) or directly from the parameter estimates of the variables. According to Hilbe (2011, p. 102), the IRR is an exponent of the model's coefficients $IRR = e^{\widehat{\beta_i}}$. This means that if the variable is positive and a dummy variable the number of trips taken increases, the IRR times for the group coded 1, compared to the group coded 0. In continuous variables, the number of trips taken increase IRR times for each unit increase of the independent variable. The direct interpretation in terms of negative binomial regression coefficients is based on the link function, here log. If the parameter estimate of a dummy variable is positive, it follows that the difference of the expected number of trips is the log amount of the parameter estimate higher for the group coded 1 compared to the group coded 0. For continuous variables, the interpretation is that if there is one unit increase in the independent variable, the expected number of trips increases by the log amount of the parameter estimate. If the parameter estimate is negative, the impact is the opposite, that is, the number of trips taken decrease.

According to Hilbe (2011, pp. 65-67), the goodness of fit tests used to assess negative binomial models are the pseudo R^2-statistics and likelihood ratio tests. The McFadden R^2 mathematical formula is $R^2 = 1 - [\ln(L_M)/\ln(L_0)]$ (Menard, 2000). Values closer to 1 indicate a better model fit, but similar to the pseudo R^2 in the logistic regression section, the interpretation is not intuitive compared to the coefficient of determination. The likelihood ratio test evaluates the difference between the baseline model and the full model. A significant test result indicates that the final model is a better fit than the baseline model.

4.3 Sample selection model

Sample selection problem

The results of the estimated travel cost model raised two issues: 1) the positive sign of the substitute interaction term, and 2) a departure from random sampling, generally called sample selection. The

latter could possibly explain the former. Non-random sampling can be due to sample design or the behavior of sampled individuals, for example non-responses in the survey (Wooldridge, 2002). Our travel cost model systematically excluded non-responses in the data, leading to the sample selection issue (Haab & McConoll, 2005, p. 203). Sample selection can be interpreted as an omitted variable bias that causes biased estimates if not treated properly (Heckman, 1977). In econometrics, an omitted independent variable leads to correlation between the variable and the error term. This correlation is called endogeneity. A situation where the information of the omitted variable cannot be observed is called unobserved heterogeneity. Sample selection issues stem from this type of situation, where endogeneity is caused by unobserved heterogeneity (Hilbe, 2011, p. 428).

The map-based questions in the survey divided the users group into two subsets. A significant portion (~750) of respondents did not respond to the map-based question, and hence to the substitute question. This split led to a classic sample selection issue, where the "conventional travel cost model" was only estimated for the portion of respondents in the users group responding to the map-based questions, causing possibly biased estimates. The reason why a portion of respondents did not respond to the map-based questions is unknown. A possible reason is that the map in the survey was difficult to use.

Econometric methods have been developed to correct sample selection bias. Heckman's (1977; 1979) two-stage method follows the basic idea that omitted observations of the dependant variable can be observed if certain selection criteria are met. Heckman-type models are composed of two parts, the selection equation and the outcome equation. Generally, the selection equation is a binary model with its own independent variables predicting the selection criteria. Haab and McConoll (2005, p. 203) have suggested that a two-stage probit-Poisson model should be used for recreation demand.

The probit-poisson model

Miranda and Rabe-Hesketh (2006) have developed a probit-Poisson sample selection model. In this case, the outcome equation models visits to the Baltic Sea and the selection equation models the probability of responding to the substitute question. The counts in the outcome equation are expected to follow the Poisson distribution. The link function for the Poisson model is again specified in the exponential or log form, as it was in the negative binomial model. The link function combines the distribution assumptions and the linear combination specifying the independent variables.

The outcome equation is defined as follows:

$$\ln(\mu_i) = x_i'\beta + u_i$$

where x_i' is the independent variable, β is the coefficient for that variable and u_i is the unobserved heterogeneity term.

The selection equation is the probit model which analyses the probability of an individual responding to the substitute question and is calculated as follows:

$$S_i^* = z_i'\gamma + v_i$$

$$S_i = \begin{cases} 1 \; if \; S_i^* > 0 \\ 0 \; otherwise \end{cases}$$

where z_i' is the independent variable, γ is the coefficient for that variable and v_i is the unobserved heterogeneity term.

Dependence between the selection equation and the outcome equation is established by the unobserved heterogeneity terms u_i and v_i. For this purpose, the unobserved heterogeneity term v_i in the selection equation is expressed as $\lambda u_i + \zeta_i$, where λ is a free parameter (a factor loading), which is estimated alongside the other parameters in the equation. ζ_i is the residual term, which has normal distribution and is independent of the unobserved heterogeneity term ε_i.

The variance of u_i determines the amount of overdispersion in the counts. Overdispersion is identified from the outcome equation and thus another parameter (sigma) is needed: $\sigma^2 = Var(\varepsilon_i)$. The total variance of the selection/outcome equation is then $\lambda^2\sigma^2 + 1$.

Dependence between the two equations is measured by the correlation of both the selection and outcome equations' unobserved heterogeneity terms. A significant correlation indicates that endogeneity is present in the outcome equation and the sample selection model is appropriate. The assumption is that the terms are jointly normally distributed. Bivariate normal distribution means that both variables follow normal distribution, independently as well as when added together.

Covariance measures the variation of two variables together, compared to variance which indicates the variation of a single variable. Here the covariance is the unstandardized relationship of the unobserved heterogeneity terms. The covariance matric of the u_i and v_i is as follows:

$$Cov[(u_i\,, v_i)'] \equiv \sum = \begin{pmatrix} \sigma(u,u) & \sigma(u,v) \\ \sigma(v,u) & \sigma(v,v) \end{pmatrix} = \begin{pmatrix} \sigma^2 & \lambda\sigma^2 \\ \lambda\sigma^2 & \lambda^2\sigma^2 + 1 \end{pmatrix}$$The strength of the relationship between the unobserved heterogeneity terms is assessed by correlation ρ. Correlation is a scaled version of covariance. Correlation is calculated as follows:

$$\rho = \frac{\lambda\sigma^2}{\sqrt{\sigma^2(\lambda^2\sigma^2 + 1)}}$$

$\rho = 0$ implies that no relationship exists between the selection and outcome equations. If this is the case, the data only modelled the outcome equation.

There are a few advantages to presenting the model in this way. The variance of the counts is identified, hence the model is not restricted by the Poisson distribution assumption of equidistribution; if $\sigma \neq 0$, the model is overdispersed. This is even the case when $\rho = 0$, because the λ, a free parameter (a factor loading) is introduced to the selection equation.

Variables

The independent variables of the outcome equation of the sample selection model are identical to the negative binomial travel cost model. The independent variables in the selection equation should explain why respondents did not respond to the substitute question. The assumed reason is that the map-based questions were too difficult for a portion of the respondents. Possibly older, less educated respondents and respondents who are not interested in recreation could have had difficulties with the map-based questions. The descriptive statistics of the sample section implied that our assumption could be correct, because at least the age and education levels differ in the subsamples.

The binary predictor variable of the selection equation is named subs_q. Respondents who responded to the substitute question coded 1 and those who did not were coded 0. The independent variables of the equation are age, education level and importance of recreation.

Similar to the process introduced in the logistic regression section, preliminary testing with Pearson's chi-squared test and a one-way ANOVA confirmed that selected variables could be significant predictors in the selection equation. The categorical age variable was transformed to a dummy variable, exploiting the ratio of the observed and expected values of Pearson's chi-squared test. The education and importance of recreation variables are identical to the variables introduced in the travel cost model. The descriptive statistics of the variables are presented in Table 4. We estimated the probit model independently of the sample selection model to confirm that the variables are significant predictors. The probit model is presented in Appendix A.

Table 4

Variables for the selection equation

Dependent variable		Mean	Std. Dev.	Min	Max	N
Subs_q	0=No response, 1=response	0,78	0,42	0	1	3395
Independent variable						

Age	0=18-64, 1=Over 65	0,14	0,34	0	1	3394
Education	0=Lower, 1=Higher	0,56	0,50	0	1	3394
Importance of recreation	0=0-25, 1=26-100	0,52	0,50	0	1	3393

Estimation and goodness of fit of the model

Miranda and Rabe-Hesketh (2006, pp. 1-5) have stated that consistent estimators for the presented sample selection model can be obtained using a maximum likelihood estimation. The likelihood of the sample selection model is marginal likelihood, where the evaluation and maximization demand that the unobserved heterogeneity term ε_i must be integrated out. The Newton-Raphson algorithm or adaptive quadrature was used for this process, because generally no closed form expressions exist for marginal likelihoods. Rabe-Hesketh, Skrondal and Pickles (2002), in their comparison of approximation methods, favour adaptive quadrature in the Poisson model setting. The Stata Manual (2017, pp. 14-24) defines the factors influencing the accuracy of the adaptive quadrature approximation and demonstrates its appropriate usage. The accuracy of the approximation method depends on three factors, namely the number of quadrature points, the location of the points and the smoothness of the approximated function. In the used sample selection model, the location of the points are automatically modified according to the distribution of ε_i (Miranda & Rabe-Hesketh, 2006, p. 5). The approximation of the log likelihood by adaptive quadrature is deemed to be good when the coefficients of the model do not change more than 0,01%. The results can confidently be interpreted when this level of accuracy is achieved. In this thesis, the appropriate number of quadrature points was 12-16, depending on the independent variables of the model. Estimation started by adding all the independent variables to the selection and outcome equations and the insignificant variables were dropped from the outcome equation by stepwise estimation.

The interpretation of the coefficients of the sample selection model's outcome equation and the Poisson model is identical to the negative binomial travel cost model. The estimated sample selection model's overall fit was tested with the Wald test. This test is equivalent to the likelihood ratio test (Hosmer, 2013, p. 14). Look for the calculation of the test statistic look (Greene, 2002, p. 487). The reported test statistic excludes all variables from the selection and outcome models, and thus the reported log likelihood is the sum of both models' likelihoods (Miranda & Rabe-Hesketh, 2006, p. 14). The significant test statistic implies that the model fits the data better than the baseline model. The final model is presented in Table 7.

4.4 Welfare estimate

Haab and McConnell (2005) have argued that the value of access to the recreation site is the only reliable welfare estimate to be gained from the single-site travel cost model. The monetary value of a

quality change of the site under valuation is difficult to calculate, because the timeframe of the data collection in the form of a survey is limited, leading to a situation where the variability of the site condition is also limited. A consumer surplus is generally deemed to be a good estimate of welfare. The consumer surplus or access value per season can be calculated by integrating the area with the expected demand function, as follows:

$$CS = \int_{C^0}^{\infty} e^{\beta_0 + \beta_{tc} C} \, dC = \frac{\lambda_i}{\hat{\beta}_{tc}}$$

where λ_i is the expected number of trips, $\hat{\beta}_{tc}$ is the parameter estimate for the travel cost variable and C^0 is the current travel cost. It is also assumed that for the exponential demand function the price where trips fall to zero, called the choke price, is infinite.

The consumer surplus value per season can be used to calculate the per-trip consumer surplus by dividing the per-season CS by the expected number of trips taken (Blaine et al., 2014; Parsons, 2003; Edwards et al., 2011), as follows:

$$CS = \frac{(\frac{\lambda_i}{\hat{\beta}_{tc}})}{\lambda_i} = \frac{1}{\hat{\beta}_{tc}}$$

In this thesis, the estimated model included the travel cost substitute interaction term. Travel cost interaction term influence the steepness of the slope of the recreation demand curve and the per-trip consumer surplus then calculated following way (Lankia et al., 2017; Loomis et al., 2001).

$$CS = \frac{1}{\hat{\beta}_{tc} + \hat{\beta}_{tc*subs}}$$

The mean seasonal access value per person is calculated by multiplying the per-trip CS by the model's predicted mean number of trips taken by the individual. The welfare estimates of recreation at the Baltic Sea are presented in Table 8.

2. Results

This chapter presents the logistic regression model, the five travel cost models and the extended travel cost model, and the sample selection model. For all of the models only significant variables are presented. The chapter concludes by discussing the welfare estimates.

5.1 Logistic regression: Who has substitute sites?

The odds ratio (OR) is the most intuitive way to interpret a logistic regression model. An example of interpreting the results of a model using the odds ratio is as follows – the outcome here refers to an

individual substitute site. If the independent variable is education, the OR is 1,660, meaning that if the individual has a university/polytechnic education, the odds of having a substitute site are 1,660 times greater than if the individual does not. Dummy variables indicating respondents' home countries (Finland and Latvia) are interpreted in a similar manner, but the odds are compared to the reference group (Germany).

According to univariate Wald tests, Finland is the only nonsignificant independent variable in the final model. All other variables are significant at a p-value <0.05. Despite the fact that Finland is non-significant, interpretation of the model demands that two out of three variables indicating respondents' home countries have to be included in the model. The third variable, Germany, was selected for the reference group.

Table 5 logistic regression model

		β	Exp(β)	p-value
Constant		-0,605	0,546	0,000
Timing of visitation	0=Over 12 months ago, 1=Inside 12 months ago	0,352	1,422	0,001
Number of activities	0=One, 1=Two or more	0,439	1,551	0,000
Number of facilities	0=None or some, 1=Many	0,190	1,209	0,036
Age	0=Between 18-29 and over 65, 1=Between 30-64	0,337	1,401	0,001
Education	0=Lower education, 1=University/polytechnic	0,507	1,660	0,000
Finland	0=Not Finnish, 1=Finnish	-0,079	0,924	0,462
Latvia	0=Not Latvian, 1=Latvian	-1,415	0,243	0,000
Number of observations		2417		
Number of respondents		2642		
Log likelihood		3056,458		
Restricted log likelihood (constant only)		3185,187		
Log-likelihood ratio		257,132		0,000
Wald test		36,798		0,000
Cox & Snell R^2		0,101		
Nagelkerke R^2		0,135		
Hosmer-Lemeshow test		31,885		0,000
Overall predictive capability %		61,605		

The likelihood ratio and the Wald test are both significant at the p-value < 0.05, hence the model at least predicts the outcome better than the baseline model. Both pseudo R^2 values are small and the Hosmer-Lemeshow test is significant at the p-value < 0.05 level, indicating that the model is not a good

fit for the data. The overall predictive capability is 61,6%, thus the model predicts moderately well the outcome variable groups of not having a substitute site and having a substitute site. All goodness of fit tests, assessed together, lead to the conclusion that the model fits the data poorly. Due to the poor predictability of the model, the interpretation focuses on the sign of the variables. A positive sign indicates higher odds in the group coded 1 and a negative sign indicates lower odds of having a substitute site.

The odds of having a substitute site increase if an individual visited within 12 months, has a university or polytechnic education and participates in more than one recreational activity at the Baltic Sea. The odds also increase if an individual is between the ages of 30 and 64, compared to the odds for younger and older individuals. If a respondent perceived that there are many compared to none/some facilities at the most frequently visited site, the odds also increase. The only variables decreasing the odds of having a substitute site are country variables, indicating that the odds of having a substitute site are smaller for Latvians and Finns compared to Germans. The length of visit, the importance of recreation, the travel mode, the number of perceived species at the most frequently visited site, income and distance from the respondents' homes to the most frequently visited recreation site were insignificant in the final model. These results will be further analysed the discussion section.

5.2 Single-site travel cost model: Demand for recreation at the Baltic Sea

All five of the travel cost models presented here have their own purpose. The first two models form the basis for the analysis and the latter three models contribute to the analysis of substitutes. The complexity of the models increases from the first to the last model.

Table 6

Travel cost model (negative binomial regression Model 1, negative binomial regression Model 2)

	Model 1 - TC		Model 2 - TC		Model 3 - TC + SUBS		Model 4 - TC + TC*SUBS		Model 5 - TC + TC*SUBS + SUBS	
	Negative binomial		Negative binomial		Negative binomial		Negative binomial		Negative binomial	
	β	p-value	β	p-value	β	p-value	β	p-value	β	p-value
Constant	2,0889	0,000	1,0506	0,000	1,1153	0,000	1,0575	0,000	1,1954	0,000
TC	-0,0033	0,000	-0,0031	0,000	-0,0040	0,000	-0,0049	0,001	-0,0054	0,000
TC*SUBS							0,0015	0,000	0,0022	0,000
Substitute					-0,1277	0,041			-0,2439	0,000
Income	0,0002	0,000	0,0001	0,003	0,0002	0,000	0,0001	0,000	0,0001	0,000
Age			0,0097	0,000	0,0064	0,003	0,0075	0,000	0,0064	0,003
Occupation			-0,1336	0,044						
Importance of recreation			0,3450	0,000	0,3469	0,000	0,3917	0,000	0,3489	0,000
Number of activities			0,6588	0,000	0,6613	0,000	0,6482	0,000	0,6576	0,000
Multi-purpose trips			-0,1411	0,017			-0,1320	0,044		
Finnish			0,4179	0,000	0,3641	0,000	0,3825	0,000	0,3701	0,000
Latvian			-0,2342	0,028	-0,3543	0,001	-0,3223	0,004	-0,3789	0,000
Alpha (α)	1,7205	0,000	1,5899	0,000	1,5490	0,000	1,5519	0,000	1,5353	
Number of observations	2388		2306		1806		1827		1806	
Number of respondents	3395		3395		3395		3395		3395	
Log likelihood	-7449,9		-7117,9		-5629,7		-5703,0		-5619,9	
Restricted log likelihood (constant only)	-7613,1		-7395,8		-5857,8		-5945,3		-5857,8	
Log-likelihood ratio	326,34	0,000	555,76	0,000	456,05	0,000	484,42	0,000	475,72	
McFadden's Pseudo R^2	0,0214		0,0380		0,0390		0,0407		0,0410	

Based on the reported low R^2 value, the models do not fit the data well, but according to the significant log-likelihood ratio test, the models are at least better than the baseline model. The significant likelihood ratio test alpha (α)=0 indicates that the data are overdispersed and the negative binomial model is the appropriate model to estimate the data, rather than the Poisson model. These interpretations apply to all five negative binomial models.

All of the estimated models have common characteristics; the independent variables and variable significance levels as well as the signs of the variables are consistent in all of the models. Model 1 is the simplest travel cost model, and includes only travel costs and income. The mean predicted number of trips taken by respondents in Model 1 is 9.37. The sample used in the estimation includes all the users. The number of observations is lower than the number of respondents because of incomplete observations. Also, all independent variables are significant at the $p<0.05$ level. Most importantly, the travel cost variable is negative in both models, leading to the ordinary downward sloping demand function for recreation. For example, the IRR for the income variable is $IRR = e^{0,0002} = 1.0002$, meaning that for one euro increase in income, the number of trips taken increases to 1.0002. The implication is that the higher the income, the more trips to the Baltic Sea an individual takes. However, the model does not fit the data well, and therefore the focus is more on the signs of the variables. A negative sign indicates a decrease in the number of trips taken and a positive sign indicates an increase.

Model 2 is an extension of Model 1, including all independent variables in the estimation but excluding both substitute variables. The number of observations is lower than the number of respondents, because of incomplete observations, and all users are included in the estimation. All independent variables are significant at the $p<0.05$ level. Again, the travel cost variable is negative, producing the ordinary downward sloping recreation demand curve. The mean predicted number of trips taken by respondents in Model 2 is 9.49.

The positive income variable indicates that visitors to the Baltic Sea with higher income levels take more trips compared to those with lower income levels. The implication of the positive age variable is identical, meaning the older the individual is, the more trips to the Baltic he/she takes. The occupation dummy variable indicates that having a high school or university education decrease the number of trips taken compared to having lower than a high school education. The importance of recreation variable has the opposite impact: if recreation is important to the individual, the number of visits to the Baltic Sea increases compared to individuals that do not consider recreation important. The number of trips also increases if an individual participates in more than one recreational activity at the Baltic Sea instead of only one, but decreases if recreation is the only purpose of the trip compared to having other purposes besides recreation. The

variables indicating the respondents' home countries are compared to the reference group of Germans. The implication is that Finns visit more and Latvians less than Germans.

For Model 3, the substitute dummy is included as an independent variable in the estimation process. All variables are significant at $p<0.05$ in the model. The number of observations is lower than for Models 1 and 2 because of incomplete observations. The substitute dummy in the model allows for calculating the predicted number of trips taken for individuals with a substitute site (9.80) and without a substitute site (9.54). The travel cost parameter is again negative, producing a downward-sloping demand function for recreation, but the occupation and purpose of the trip variables are no longer significant. All other variables have an identical impact on the number of trips taken as in Model 2. The substitute dummy is negative, implying that the number of trips taken decreases if an individual has substitute site for the Baltic Sea compared to those without substitute sites.

In Model 4, the substitute dummy is omitted from the analysis and the substitute travel cost interaction term is introduced to the analysis. Similar to Model 3, this model is also impacted by sample selection and incomplete observations. The number of observations is lower than the number of respondents. All independent variables are significant at the $p<0.05$ level. Compared to Model 2, individuals' occupation is no longer a significant predictor and does not impact the number of trips taken to the Baltic Sea, but the purpose of the trip is a significant predictor. All significant variables have an identical impact on the number of trips taken, as in Model 2. According to Model 4, the predicted number of trips taken is 9.88.

Due to the interaction term, Model 4 produces two distinct recreation demand curves, one for individuals that have a substitute site for the Baltic Sea and one for individuals that do not. The slope of the demand curve for individuals without substitute sites is the parameter estimate for the variable TC, and for individuals with substitute sites it is TC+TC*SUBS. Contrary to expectations, the sign of the interaction term is positive, leading to a less steep demand function for those who have substitute sites compared to those visitors without substitute sites.

Model 5 includes a substitute site dummy and a substitute interaction term and thus all variables are included in the estimation. The travel cost substitute term is again negative, leading to downward-sloping recreation demand curves. All independent variables are significant at the $p<0.05$ level and the impact of the independent variables is identical to the other four models. On the one hand, a negative substitute dummy indicates that if an individual has a substitute site, he/she takes fewer trips to the Baltic Sea. On the other hand, a positive substitute travel cost interaction term indicates that the slope of the recreation demand curve is less steep compared to individuals without substitute sites. The predicted number of trips taken by individuals with substitute sites is 9.71 and by individuals without substitute sites 9.58.

Most of the demographics, the site condition variable and the travel mode variable were insignificant in all models. These variables do not affect the number of trips taken to the Baltic Sea. A more detailed interpretation of the results and their meaning is provided in the discussion section.

5.3 Sample selection: an extended single-site travel cost model

Due to the unintentional exclusion of a portion of the respondents, namely those who did not reply to the map-based questions, the possibility of sample selection issues exists. The estimated probit-poisson sample selection model is presented here. Contrary to the negative binomial travel cost models, all of the 3,395 users are included in the estimation. The model is estimated based on variables of travel cost Model 4. The substitute dummy variable is excluded, and substitute interaction term is included. Added to this, the variables in the selection equation are excluded from the estimation of the outcome equation.

Table 7
Travel cost model (probit-poisson sample selection Model 3)

Outcome equation		
	B	p-value
Constant	0,6237	0,000
TC	-0,0037	0,000
TC*SUBS	0,0013	0,007
Income	0,0002	0,000
Finnish	0,4894	0,000
Latvian	0,2109	0,060
Time	-0,0009	0,040
Number of activities	0,5664	0,000
Number of facilities	-0,1470	0,025
Selection equation		
Constant	0,5642	0,000
Education	0,2091	0,000
Importance of recreation	0,2431	0,000
Age	-0,2219	0,001
Rho (ρ)	-0,1518	0,066
Sigma (σ)	1,1290	0,000
Likelihood ratio test for rho (ρ)=0	37323,27	0,000
Number of observations	3395	
Number of respondents	3395	
Log likelihood	-6681,29	
Restricted log likelihood (constant only)	-25342,9	
Wald chi2	304,98	0,000

Rho(ρ) measures the correlation between the unobserved heterogeneity terms. The correlation is significant at level 0.1, but the likelihood ratio test Rho (ρ)=0 is significant at level 0.05. The tests indicate that there is a dependency between the selection equation and the outcome equation. The interpretation is that the

model corrects the portion of the selection bias caused by the excluded group of respondents earlier labelled "no substitute question" and the use of a sample selection model is justified. The significant boundary likelihood ratio test sigma σ=0 indicates that the distribution of trips is overdispersed. An identical result is reported by the negative binomial travel cost models. According to the significant Wald test, the sample selection model fits the data better than the baseline model.

In results of the negative binomial travel cost model, the pseudo coefficient of determination McFadden R^2 is reported. McFadden R^2 is based on the ratio of log-likelihoods between the baseline model (intercept only) and full model. Calculus of the pseudo R^2 is possible for the sample selection model, but interpretation of its meaning would be unclear, at least it would not have original meaning of McFadden R^2 because the reported log-likelihoods are sum of the selection - and outcome equation. We decided not to report it.

In the selection equation, all variables are significant at the p<0.05 level, indicating that education, importance of recreation and age predict the probability of replying to the substitute question. In the outcome equation, all other variables except the variable Latvian are significant at the p<0.05 level. The reason to include country variables in the sample selection model is identical to that of the logistic regression and negative binomial regression. It is necessary to have two out of the three for the interpretation to be possible.

The interpretation of the outcome model is identical to the negative binomial model. Most importantly, the travel cost parameter is negative and significant, indicating that with higher travel costs fewer trips are taken to the Baltic Sea, and the demand curve for recreation is downward-sloping, as expected. Income has a positive impact on the number of trips taken. The higher the individual's income is, the more visits to the Baltic Sea are taken. Individuals earning a higher income tend to visit the Baltic Sea more often. Both country variables are positive, indicating that Latvians and Fins take more trips to the Baltic Sea compared to Germans.

The time variable is negative, as expected: the longer individuals tend to stay at the Baltic Sea, the fewer the number of trips that are taken. The number of activities variable is positive. Individuals who participate in more than one recreational activity at the Baltic Sea take more trips than those who participate in only one recreational activity. The number of facilities variable is negative, indicating that if an individual perceived that the most frequently visited site had many facilities, fewer trips are taken.

The variables indication the gender of the respondent, the occupational status, the household size, the perceived water clarity of the most frequently visited site, the travel mode used to arrive at the most frequently visited site, and whether recreation was the only purpose of the trip were insignificant in the outcome equation.

5.4 Welfare estimates of recreation at the Baltic Sea

A good estimate of the welfare of recreation is consumer surplus. The formula for the per-trip consumer surplus (CS) as well as the annual CS estimates were presented earlier, and the consumer surplus was calculated by using the demand functions from the five travel cost models and the sample selection model.

Table 8
Consumer surplus estimates for recreation at the Baltic Sea

	CS per trip (€)	95% CI	Annual CS estimates (€)
Model 1 (TC)	301,75	[301.75, 301.76]	2828,5
Model 2 (TC)	325,23	[325.23, 325.25]	3086,4
Model 3 (TC)	250,68	[250.68, 250.69]	2420,5
Model 4 (TC)	202,38	[202.37, 202.40]	1835,4
Model 4 (TC + TC*SUBS)	290,88	[290.87, 290.90]	3188,9
Model 5 (TC)	186,06	[186.05, 186.07]	1782,0
Model 5 (TC + TC*SUBS)	316,64	[316.62, 316.65]	3074,2
Model 6 (TC)	273,40	[273.37, 273.42]	*3563,2
Model 6 (TC + TC*SUBS)	420,50	[420.46, 420.54]	*5545,6

Confidence intervals for the per-trip consumer surplus are calculated using the Taylor approximation (Loomis et al., 2001; Hesseln et al., 2004)
* Model 6 values are sample means of visits to the Baltic Sea, not the model's predicted values.

Models 1 and 2 produce welfare estimates of recreation at the Baltic Sea for all users. The estimates of the welfare of a single trip to the Baltic Sea are 301.75€ and €325.23€, respectively, based on these models. Models 3, 4 and 5 capture the impact of substitutes on the welfare estimates, but sample selection causes bias to the estimates and the estimates from the models are only applicable to a subsample of users.

Model 3 introduced a substitute dummy to the estimation process. The substitute dummy indicates whether an individual has a substitute site for recreation at the Baltic Sea or not. As expected, introducing the variable to the analysis lowers the demand for recreation and the welfare estimate for a single trip to the Baltic Sea for all users is lower compared to Models 1 and 2. The per-trip welfare estimate according to Model 3 is 250.68€.

Model 4 omits the substitute dummy from the analysis but introduces the travel cost substitute interaction term to the analysis. The interaction term indicates the sensitivity of respondents with substitute sites to the demand for recreation at the Baltic Sea, in other words, how much an increase to the travel costs changes the number of trips taken to the Baltic Sea. Due to the interaction term, two welfare estimates are obtained from Model 4: 1) one for individuals with substitute sites at 290.88€, and 2) one for individuals without substitute sites at 202.38€. The higher welfare estimate for individuals with substitute sites is due to a more elastic demand curve compared to individuals without substitute sites.

Model 5 includes both the substitute dummy and the travel cost interaction term in the analysis. Similar to Model 4, two welfare estimates are calculated: 1) for individuals with substitute sites at 316,64€, and 2) for individuals without substitute sites at 186.06€. The impact of the substitute dummy on these estimates is low. The sample selection model produces higher welfare estimates compared to Models 4 and 5. The welfare estimates of the sample selection model are 273.40€ and 420.50€, respectively. All of these results will be further analyzed in the discussion section.

3. Discussion

This thesis answered two questions, namely who had substitute sites to the Baltic Sea, and how substitute sites impact recreation at the Baltic Sea. We estimated the logistic regression model to answer the former question and the negative binomial travel cost model to value recreation at the Baltic Sea. Unexpected results for the negative binomial travel cost model caused us to suspect sample selection issues in the data. We estimated the travel cost model with the probit-poisson sample selection model to address the issue.

Logistic regression

Awareness of possibilities and the different constraints individuals encounter influence whether an individual has substitute site or not. These factors probably also explain the unintuitive sign of the site condition variable and the number of facilities variable. Individuals perceiving to have many facilities at their most frequently visited site have higher odds of having a substitute site than those perceiving to have none or some. The former are likely to be more aware of different sites and possibilities than the individuals visiting sites with fewer facilities. It is also possible that this variable is a poor indicator of site condition, for example sites with many facilities could also be crowded. Even though awareness of possibilities increases as age increases, the different constraints experienced also change over time. Constraints probably explain the higher odds of individuals between the ages of 30 and 64 having substitute site compared to the odds of younger and older individuals having substitute site. It is likely that constraints limit older and younger individuals more than those in the middle, for example some sites could be difficult to access by older people, whereas younger people tend to have less money to use for recreation. The higher odds of having a substitute site among individuals with a higher education, those visiting within 12 months, and those participating in more than one recreational activity at the Baltic Sea are explained by the greater awareness of possibilities and more interest in recreation relative to less educated individuals, non-frequent visitors and individuals participating in only one recreational activity.

The negative sign of the variable Finland and the non-significance of the distance from home to the recreation site in the final model is unexpected. Despite the negative sign, it should be remembered that the odds of individuals from Finland compared to the odds of Germans are really close to one another, indicating that the odds of Fins having substitute site are nearly equal to the odds of Germans having substitute sites.

The non-significant distance variables in the logistic regression analysis were more concerning, because distance has great importance in environmental valuation. The fundamental assumption in recreation valuation is that, as distance and travel time to the site increase, travel costs also increase. It is also clear that as distance increases, the number of possible substitutes increases.

We had two separate distance variables, the Euclidean distance from the respondents' homes to the recreation site and the Euclidean distance from their homes to the Baltic Sea coast. Only the distance from respondents' homes to the recreation site was significant according to preliminary testing, but the variable is non-significant in the final model. Even with the non-significant result in the preliminary testing, estimation was also performed with the distance from respondents' homes to the Baltic Sea coast, to highlight the importance of distance as an explanatory factor. The variable was significant in the models, but the results were counterintuitive; as distance increased, the odds of having substitute sites decreased.

There are possibly two explanations for the distance effects in the logistic regression. Parsons (1991) has suggested that individuals favoring recreation already reside closer to recreation sites. In a CE study, Jörgensen et al. (2012) found that users live closer to substitute sites than non-users. It could be that individuals valuing recreation highly live close to the Baltic Sea and also have substitute sites, diminishing the distance effects in the logistic regression model. This argument is not supported by our preliminary testing, because the mean points indicated that the importance of recreation is higher among respondents who do not have substitute site (see Appendix A for the difference). The second possible explanation is sample selection. It is possible that the portion of users excluded from the analysis could have significantly different residential locations relative to the Baltic Sea, as well as to the substitute sites, compared to the subsample under analysis. We tried to estimate the sample selection model, but it failed to converge. Even with the significant distance from respondents' homes to the Baltic Sea coast in the logistic regression model, we decided to present the model without the distance variable. The presented model was more consistent with the estimation process.

Travel cost model and sample selection model

In the negative binomial travel cost models, the significant variables were respondents' home countries, income, age and, in Model 2, individuals' occupation. Other significant predictors were the number of recreational activities in which respondents participated, the importance of recreation and, in Models 2 and 4, the purpose of the trip. All of these variables have the expected impact on the number of trips taken. The impact of substitutes is captured in Models 3, 4, 5 and the sample selection Model 6. The mean of the predicted number of trips varies between ∼9.07 and 10.96, depending on the model. This is lower than the observed mean of the data of 13.09.

The substitute dummy variable indicated that individuals with substitute take fewer trips to the Baltic Sea, but the positive effect of the substitute interaction term on the slope of the recreation demand is unexpected. A positive interaction term leads to a more inelastic demand curve for those with substitute compared to those without substitute, meaning that increasing the travel costs does not influence the number of visits among individuals with substitute as much as it influences individuals without substitute. According to expectations, the impact should be the opposite: as travel costs increase, people with substitute should change their recreation site to the substitute site.

There are two possible explanations for the counterintuitive results: 1) it could be that people with a high preference for recreation at the Baltic Sea also have substitute, and 2) the sample selection, as it is possible that the omitted information due to the excluded group of respondents led to the unexpected result.

The presented sample selection model was estimated with identical independent variables to that of the negative binomial Model 4. Added to this, the sample selection model was also estimated with identical variables to Model 5, but the results were unreliable. None of the substitute variables were not significant in the model. This is probably due to the correlation with the substitute dummy and the interaction variable. This correlation had only a minor effect on the negative binomial Model 5; the magnitude of the estimated parameters diverged only slightly from the other negative binomial models.

The sample selection model confirmed the existence of a sample selection bias and probably corrected a share of the bias present in the data, leading to different significant predictors compared to the negative binomial travel cost models. The independent variables importance of recreation, age and purpose of trip no longer influenced the number of trips taken to the Baltic Sea. The variable Latvian is positive compared to the negative binomial travel cost models, implying that the number of visits to the Baltic Sea is higher among Latvians than among Germans. The reason for the variation is the increased number of respondents included in the model. Interestingly, the facilities variable, indicating the site condition, is significant in the sample selection model. The negative impact of the variable on the number of trips taken is unexpected, implying that individuals perceiving to have many facilities at their most frequently visited site take fewer visits to the Baltic Sea. A possible explanation for the results is identical with that of the logistic regression model; it could be that the variable is not a good indicator of site condition.

The counterintuitive sign of the substitute travel cost interaction term still remains in the sample selection model, implying that the explanation for the higher demand among individuals with substitutes is due to the greater commitment and interest in recreation, not the sample selection.

Welfare estimates and limitations of the study

The welfare estimates derived from the travel cost models vary between ~186€ and 420€. The confidence intervals for the welfare estimates are negligible in all of the models. The explanation for this is the minor variances of the original TC and TC*SUBS parameter estimates, as variances are used to calculate the confidence intervals.

Based on the Model 2 per-trip consumer surplus, the estimate for recreation at the Baltic Sea is 325.23€. It was expected that this estimate, capturing all users, should be between the consumer surplus estimates from Models 4 and 5, producing the welfare estimates for users with substitutes and those without. However, the estimate from Model 2 is higher than the estimates from Models 4 and 5. The reason for the different magnitude is sample selection. Furthermore, the welfare estimate of Model 3 captures the impact of substitute on the number of trips taken, leading to a lower welfare estimate, probably amplified by sample selection, and the true impact of substitute is lower.

The welfare estimates from the sample selection model are higher compared to the negative binomial Models 3, 4 and 5, but the welfare estimate of Model 2 is between the welfare estimates obtained from the sample selection model. This is reasonable, because Models 1, 2 and the sample selection Model 6 are estimated for all of users, but Models 3, 4 and 5 are affected by sample selection, leading to more consistent estimates from Models 1, 2 and 6 compared to Models 3, 4 and 5.

There are only a limited number of revealed preference studies focusing on the recreational value of the Baltic Sea. Czaikowski et al. (2015) studied value of recreation at the Baltic Sea in nine countries using the TC method, arriving at more conservative values for recreation at the Baltic Sea. Their country level estimates are: for Germany 31.4763€, for Finland 80.6823€ and for Latvia 28.3449€, which are significantly lower than our estimates.

There are three separate differences compared to the study by Czaikowski et al. (2015) that likely inflate the welfare estimates of this thesis: the econometric model, modelling and data: 1) the negative binomial model vs. the zero-inflated negative binomial model, 2) excluding non-users from the analysis, and 3) having a smaller number of non-users in the data. Furthermore, the unique distinction of the earlier study is the sample selection issue, which caused bias to the logistic regression and the negative binomial model. The sample selection model probably captured a portion of this sample selection bias, but the majority of welfare estimates remain large.

We also had modelling differences: this thesis used a different variant of negative binomial distribution. This distribution assumes there are zero observations in the data, but we omitted zero observations from the dependant variable. In a situation where the entire distribution is not observed, the appropriate probability

density distribution to model the data is truncated negative binomial, namely the zero-truncated negative binomial distribution. For example, Grogger and Carson (1991) demonstrated that parameters are biased if truncated data are modelled with the conventional negative binomial model. We estimated the zero-truncated negative binomial model, but decided not to report it for two reasons. First, the alpha (α) values indicating overdispersion in the data were unusually high. We tried to even out the distribution of the trips by omitting outliers, but the high alpha parameter remained. Second, the probit-poisson sample selection model could not account for the truncation, thus with the conventional negative binomial distribution our results from the two models are more comparable.

The econometric model differences of our travel cost parameter are relatively simple, excluding the travel time and the weighted approach to adjust to welfare estimates to account for multi-purpose trips. However, multi-purpose trips are captured as a separate dummy variable in the analysis. We justified keeping the travel cost parameter as simple as possible based on the objective of this thesis, namely to highlight the impact of substitutes on the welfare as well as number of trips taken. The simplicity of the travel cost variable likely inflates the welfare estimates because they are obtained directly from the variable.

We could have added substitutes to the analysis as a travel costs to the substitute site. Instead of this usual approach we used dummy variable and interaction variable. The travel costs substitute captures the impact of substitute sites on the number of trips taken, as the travel cost to the substitute site increases, the number of trips to the site decreases. However, the direct impact of the substitute site demand curve would have been omitted from the analysis.

7. Conclusions

The Baltic Sea is an important recreation destination for the residents of the coastal countries. The impact of substitute on the welfare of as well as the demand for recreation has been ignored in earlier valuation research related to the Baltic Sea and addressed relatively simply or not at all in Finnish recreational valuation studies. This thesis has added to the earlier environmental valuation literature on the Baltic Sea by specially focusing on substitutes and the impact of substitutes on the welfare estimates, using the single-site travel cost method. The estimated welfare of a single recreational visit to the Baltic Sea ranges from 180€ to 420€.

We used the substitute site dummy and the travel cost substitute site interaction term to assess the impact of substitutes on the number of trips taken to the Baltic Sea, as well as the impact of substitutes on the welfare estimates. As expected, having a substitute site decreased the number of trips taken, but, contrary to expectations, the substitute site travel cost interaction term was positive, leading to a higher welfare estimate for a single trip to the Baltic Sea among those individuals with substitute site compared to individuals without substitute site. It seems that individuals with substitutes take fewer visits to the Baltic

Sea, but their demand for recreation is more inelastic compared to individuals without substitutes. It was also found that demographics, site condition, actual timing of visitation and the number of activities in which respondents participate explain who has substitute sites for the Baltic Sea.

The approach of allowing respondents to define the substitute sites could solve the unique problem of determining the relevant substitute sites caused by everyman's rights in the Nordic countries. Map-based questions seem a reasonable method for this, but further development of the questions is required to avoid non-responses, leading to sample selection issues. Future research could possibly use the information acquired in this thesis about the location and features of substitute sites for random utility travel cost modelling.

It is difficult to draw very direct policy implications from the thesis results. Neither logistic regression nor travel cost models produced clear predicting variables that could aid decision-making. However, it can be said that awareness of possibilities increases the possibility of having a substitute site, while informing citizens about different possibilities could avoid a significant loss of welfare in situations where recreation at the preferred site is not possible.

References

Akron, A., Ghermandi, A., Dayan, T. & Hershkovitz, Y. 2017, Interbasin water transfer for the rehabilitation of a transboundary Mediterranean stream: An economic analysis, *Journal of environmental management,* **202**, s. 276-286.

Alves, B., Ballester, R., Rigall-I-Torrent, R., Ferreira, Ó & Benavente, J. 2017, How feasible is coastal management? A social benefit analysis of a coastal destination in SW Spain, *Tourism Management,* **60**, s. 188-200.

Bhat, M.G. 2003, Application of non-market valuation to the Florida Keys marine reserve management, *Journal of environmental management,* **67**(4), s. 315-325.

Bin, O., Landry, C.E., Ellis, C.L. & Vogelsong, H. 2005, Some consumer surplus estimates for North Carolina beaches, *Marine Resource Economics,* **20**(2), s. 145-161.

Blaine, T.W., Lichtkoppler, F.R., Bader, T.J., Hartman, T.J. & Lucente, J.E. 2015, An examination of sources of sensitivity of consumer surplus estimates in travel cost models, *Journal of environmental management,* **151**, s. 427-436.

Boyer, T.A., Melstrom, R.T. & Sanders, L.D. 2017, Effects of climate variation and water levels on reservoir recreation, *Lake and Reservoir Management,* **33**(3), s. 223-233.

Central Statistical Bureau of Latvia. 2018, Statistics in Brief. Lāčplēša iela 1, Rīga, LV-1301, Latvija.

Czajkowski, M., Ahtiainen, H., Artell, J., Budziński, W., Hasler, B., Hasselström, L., Meyerhoff, J., Nõmmann, T., Semeniene, D. & Söderqvist, T. 2015, Valuing the commons: An international study on the recreational benefits of the Baltic Sea, *Journal of environmental management,* **156**, s. 209-217.

De Valck, J., Broekx, S., Liekens, I., Aertsens, J. & Vranken, L. 2017, Testing the influence of substitute sites in nature valuation by using spatial discounting factors, *Environmental and Resource Economics,* **66**(1), s. 17-43.

DesStat. 2019, Facts Figures. *referred: [1.4.2019]*. Access method: https://www.destatis.de/EN/FactsFigures/SocietyState/Population/CurrentPopulation/CurrentPopulation.html

Ditton, R.B. & Sutton, S.G. 2004, Substitutability in recreational fishing, *Human Dimensions of Wildlife,* **9**(2), s. 87-102.

Du Preez, M., Dicken, M. & Hosking, S.G. 2012, The value of Tiger Shark diving within the Aliwal Shoal marine protected area: a travel cost analysis, *South African Journal of Economics,* **80**(3), s. 387-399.

Du Preez, M. & Hosking, S.G. 2011, The value of the trout fishery at Rhodes, North Eastern Cape, South Africa: a travel cost analysis using count data models, *Journal of Environmental Planning and Management,* **54**(2), s. 267-282.

Edwards, P.E., Parsons, G.R. & Myers, K.H. 2011, The economic value of viewing migratory shorebirds on the Delaware Bay: an application of the single site travel cost model using on-site data, *Human dimensions of wildlife,* **16**(6), s. 435-444.

European Commission. 2018, How much biodiversity is in Natura 2000? The 'umbrella effect' of the European Natura 2000 protected area network: final report, *Office for Official Publications of the European Communities.*

European Environment Agency (EEA). 2019a, CORINE Land Cover. *European Union, Copernicus Land Monitoring Service European Environment Agency (EEA). [referred: 9.1.2019]*. Access method: https://land.copernicus.eu/pan-european/corine-land-cover

European Environment Agency (EEA). 2019b, Natura 2000 data - the European network of protected sites. *European Environment Agency (EEA) [referred: 9.1.2019]*. Access method: https://land.copernicus.eu/pan-european/corine-land-cover/clc-2012

Eurostat, 2017. Population by educational attainment level, sex and age. *referred: [1.4.2019]*. Access method: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=edat_lfs_9903&lang=en

Freeman III, A.M., Herriges, J.A. & Kling, C.L. 2014, *The measurement of environmental and resource values: theory and methods*, Routledge.

Flanders Marine Institute. 2019, Maritime Boundaries Geodatabase: Maritime Boundaries and Exclusive Economic Zones (200NM). *version 10. [referred: 9.1.2019].* Access method: http:// www.marineregions.org/

Gentner, B. & Sutton, S. 2008, Substitution in recreational fishing, *Global challenges in recreational fisheries,* , s. 150-169.

Ghermandi, A. 2015, Benefits of coastal recreation in Europe: identifying trade-offs and priority regions for sustainable management, *Journal of environmental management,* **152**, s. 218-229.

Greene, W.H. 2003, *Econometric Analysis*, Prentice Hall.

Grogger, J.T. & Carson, R.T. 1991, Models for truncated counts, *Journal of Applied Econometrics,* **6**(3), s. 225-238.

Haab, T.C. & Hicks, R.L. 1997, Accounting for choice set endogeneity in random utility models of recreation demand, *Journal of Environmental Economics and Management,* **34**(2), s. 127-147.

Haab, T.C. & McConnell, K.E. 2005, *Valuing Environmental and Natural Resources: The Econometrics of Non-Market Valuation*, Edward Elgar Publishing.

Hagerty, D. & Moeltner, K. 2005, Specification of driving costs in models of recreation demand, *Land Economics,* **81**(1), s. 127-143.

Han, J.H., Noh, E.J. & Oh, C. 2015, Applying the concept of site substitution to coastal tourism, *Tourism Geographies,* **17**(3), s. 370-384.

Hanauer, M.M. & Reid, J. 2017, Valuing urban open space using the travel-cost method and the implications of measurement error, *Journal of environmental management,* **198**, s. 50-65.

Heckman, J.J. 1977, No title, *Sample selection bias as a specification error (with an application to the estimation of labor supply functions),* .

Heckman, J.J. 1976, The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, *Annals of Economic and Social Measurement, Volume 5, number 4*, NBER, s. 475-492.

Helcom, 2018. State of the Baltic Sea – Second Helcom holistic assessment 2011-2016. Baltic Sea Environment Proceedings 155.

Hellerstein, D.M. 1991, Using count data models in travel cost analysis with aggregate data, *American Journal of Agricultural Economics,* **73**(3), s. 860-866.

Hicks, R.L. & Strand, I.E. 2000, The extent of information: its relevance for random utility models, *Land Economics,* **76**(3), s. 374-385.

Hilbe, J.M. 2011, *Negative binomial regression*, Cambridge University Press.

Hosmer Jr, D.W., Lemeshow, S. & Sturdivant, R.X. 2013, *Applied logistic regression*, John Wiley & Sons.

Huhtala, A. & Lankia, T. 2012, Valuation of trips to second homes: do environmental attributes matter? *Journal of Environmental Planning and Management,* **55**(6), s. 733-752.

Hynes, S., Gaeven, R. & O'Reilly, P. 2017, Estimating a total demand function for sea angling pursuits, *Ecological Economics,* **134**, s. 73-81.

Hynes, S., O'Reilly, P. & Corless, R. 2015, An on-site versus a household survey approach to modelling the demand for recreational angling: Do welfare estimates differ? *Ecosystem services,* **16**, s. 136-145.

Jones, C.A. & Lupi, F. 1999, The effect of modeling substitute activities on recreational benefit estimates, *Marine Resource Economics,* **14**(4), s. 357-374.

Jørgensen, S.L., Olsen, S.B., Ladenburg, J., Martinsen, L., Svenningsen, S.R. & Hasler, B. 2013, Spatially induced disparities in users' and non-users' WTP for water quality improvements—Testing the effect of multiple substitutes and distance decay, *Ecological Economics,* **92**, s. 58-66.

Lankia, T., Neuvonen, M. & Pouta, E. 2017, Effects of water quality changes on the recreation benefits of swimming in Finland: Combined travel cost and contingent behavior model, *Water Resources and Economics,* s. 1-11.

Latinopoulos, D. 2014, The impact of economic recession on outdoor recreation demand: an application of the travel cost method in Greece, *Journal of Environmental Planning and Management,* **57**(2), s. 254-272.

Lizin, S., Brouwer, R., Liekens, I. & Broeckx, S. 2016, Accounting for substitution and spatial heterogeneity in a labelled choice experiment, *Journal of environmental management,* **181**, s. 289-297.

Loomis, J., Gonzalez-Caban, A. & Englin, J. 2001, Testing for differential effects of forest fires on hiking and mountain biking demand and benefits, *Journal of Agricultural and Resource Economics,* **26**(2), s. 508-522.

Loomis, J., Yorizane, S. & Larson, D. 2000, Testing significance of multi-destination and multi-purpose trip effects in a travel cost method demand model for whale watching trips, *Agricultural and Resource Economics Review,* **29**(2), s. 183-191.

Mangan, T., Brouwer, R., Lohano, H.D. & Nangraj, G.M. 2013, Estimating the recreational value of Pakistan's largest freshwater lake to support sustainable tourism management using a travel cost model, *Journal of Sustainable Tourism,* **21**(3), s. 473-486.

Martínez-Espiñeira, R. & Amoako-Tuffour, J. 2009, Multi-destination and multi-purpose trip effects in the analysis of the demand for trips to a remote recreational site, *Environmental management,* **43**(6), s. 1146-1161.

Menard, S. 2000, Coefficients of determination for multiple logistic regression analysis, *The American Statistician,* **54**(1), s. 17-24.

Menard, S. & Menard, S.W. 2010, *Logistic regression: From introductory to advanced concepts and applications*, Sage.

Miranda, A. & Rabe-Hesketh, S. 2006, Maximum likelihood estimation of endogenous switching and sample selection models for binary, ordinal, and count variables, *The stata journal,* **6**(3), s. 285-308.

OECD. 2016a, Family Database. *referred: [1.4.2019]*. Access method: http://www.oecd.org/els/family/database.htm

Oh, C., Sutton, S.G. & Sorice, M.G. 2013, Assessing the role of recreation specialization in fishing site substitution, *Leisure Sciences,* **35**(3), s. 256-272.

Ovaskainen, V., Neuvonen, M. & Pouta, E. 2012, Modelling recreation demand with respondent-reported driving cost and stated cost of travel time: A Finnish case, *Journal of Forest Economics,* **18**(4), s. 303-317.

Parsons, G.R. 2003, The travel cost model, *A primer on nonmarket valuation*, Springer, s. 269-329.

Parsons, G.R. 1991, A note on choice of residential location in travel cost demand models, *Land Economics,* **67**(3), s. 360-364.

Parsons, G.R. & Hauber, A.B. 1998, Spatial boundaries and choice set definition in a random utility model of recreation demand, *Land Economics,* **74**(1), s. 32-48.

Parsons, G.R., Plantinga, A.J. & Boyle, K.J. 2000, Narrow choice sets in a random utility model of recreation demand, *Land Economics,* **76**(1), s. 86-99.

Parsons, G.R. & Wilson, A.J. 1997, Incidental and joint consumption in recreation demand, *Agricultural and Resource Economics Review,* **26**(1), s. 1-6.

Pate, J. & Loomis, J. 1997, The effect of distance on willingness to pay values: a case study of wetlands and salmon in California, *Ecological Economics,* **20**(3), s. 199-207.

Peters, T., Adamowicz, W.L. & Boxall, P.C. 1995, Influence of choice set considerations in modeling the benefits from improved water quality, *Water Resources Research,* **31**(7), s. 1781-1787.

Press, S. 2009, Stata longitudinal-data/panel-data reference manual: Release 11, .

Rabe-Hesketh, S., Skrondal, A. & Pickles, A. 2002, Reliable estimation of generalized linear mixed models using adaptive quadrature, *The Stata Journal,* **2**(1), s. 1-21.

Reynaud, A. & Lanzanova, D. 2017, A global meta-analysis of the value of ecosystem services provided by lakes, *Ecological Economics,* **137**, s. 184-194.

Rolfe, J. & Windle, J. 2012, Distance decay functions for iconic assets: assessing national values to protect the health of the Great Barrier Reef in Australia, *Environmental and Resource Economics,* **53**(3), s. 347-365.

Rosenthal, D.H. 1987, The necessity for substitute prices in recreation demand analyses, *American Journal of Agricultural Economics,* , s. 828-837.

Schaafsma, M. 2011, Spatial effects in stated preference studies for environmental valuation, .

Schaafsma, M., Brouwer, R., Gilbert, A., Van Den Bergh, J. & Wagtendonk, A. 2013, Estimation of distance-decay functions to account for substitution and spatial heterogeneity in stated preference research, *Land Economics,* **89**(3), s. 514-537.

Shrestha, R.K. & Loomis, J.B. 2001, Testing a meta-analysis model for benefit transfer in international outdoor recreation, *Ecological Economics,* **39**(1), s. 67-83.

Official Statistics of Finland. 2019, Population structure [e-publication]. ISSN=1797-5379. *Helsinki: Statistics Finland [referred: 9.1.2019]*. Access method: http://www.stat.fi/til/vaerak/index.html

OECD, 2018. OECD Employment Outlook. *referred: [1.4.2019]*. Access method: https://data.oecd.org/emp/employment-rate.htm

OECD, 2016b, OECD Better Life Index. *referred: [1.4.2019]*. Access method: http://www.oecdbetterlifeindex.org/topics/income/

Sutton, S.G. 2007, Constraints on recreational fishing participation in Queensland, Australia, *Fisheries,* **32**(2), s. 73-83.

Sutton, S.G. & Ditton, R.B. 2005, The substitutability of one type of fishing for another, *North American Journal of Fisheries Management,* **25**(2), s. 536-546.

Sutton, S.G. & Oh, C. 2015, How do recreationists make activity substitution decisions? A case of recreational fishing, *Leisure Sciences,* **37**(4), s. 332-353.

Walsh, R.G., Johnson, D.M. & McKean, J.R. 1992, Benefit transfer of outdoor recreation demand studies, 1968–1988, *Water Resources Research,* **28**(3), s. 707-713.

Ward, F.A. & Beal, D. 2000, *Valuing nature with travel cost models*, Edward Elgar Publishing.

Winkelmann, R. 2008, *Econometric analysis of count data*, Springer Science & Business Media.

Wooldridge, J.M. 2015, *Introductory econometrics: A modern approach*, Nelson Education.

World Bank. 2017, World Bank Open Data: Population, female. *referred: [1.4.2019]*. Access method: https://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS?locations=FI-DE-LV

Zhang, F., Wang, X.H., Nunes, P.A. & Ma, C. 2015, The recreational value of gold coast beaches, Australia: An application of the travel cost method, *Ecosystem Services,* **11**, s. 106-114.

Appendix A

Example question from the survey:

Please think now about all your recreation visits to the Baltic Sea, whether they took place in your own country or in the other coastal countries. We are particularly interested in your outdoor recreation visits the Baltic Sea, such as being on the beach, swimming and boating.

Q3. When was the last time you spent leisure time at the Baltic Sea or its coast?

- ○ In the last 12 months → go to Q4_a
- ○ In the last 3 years but not in the last 12 months → go to Q4_b
- ○ More than 3 years ago → skip to non-user version [Your recreation sites]
- ○ I have never visited the Baltic Sea or its coast for leisure → skip to non-user version [Your recreation sites]

Q4_a. How many times did you visit the Baltic Sea or its coast in the last 12 months to spend leisure time there? Please provide a rough estimate if you cannot remember the exact number.

_____

This link helps you calculate the number of visits.

Q5. Please mark an approximate location of your place of residence on the map provided below. This will help you to locate the recreation areas around you and researchers to calculate distances to recreation sites.
Instead of exact location, you can mark a street or an intersection that is close to your place of residence. If you have several places of residence, please select the one you spend the most time at.

Q9. What do you typically do at this site? Please choose 1-3 most important activities.

○ Walking
○ Jogging or running
○ Swimming
○ Fishing
○ Nature-watching
○ Diving or snorkelling
○ Boating or sailing
○ Sun-bathing
○ Picnicking or being on the beach
○ Being at the summer house
○ Being on a cruise
○ Going to the sauna
○ Ice-skating
○ Skiing on the ice
○ Other, please specify: _____

Q15. What were your approximate one-way travel costs on your trip to the site? Please report your share of the travel costs of a one-way trip in euros, e.g. fuel and public transportation tickets (excluding meals and accommodation).

_____

Q16. How important was recreation at the sea as the purpose of the trip?

- the only purpose of the trip
- more important than other purposes, but it was not the only purpose
- equally important as other purposes
- less important than other purposes
- only a small purpose of the trip

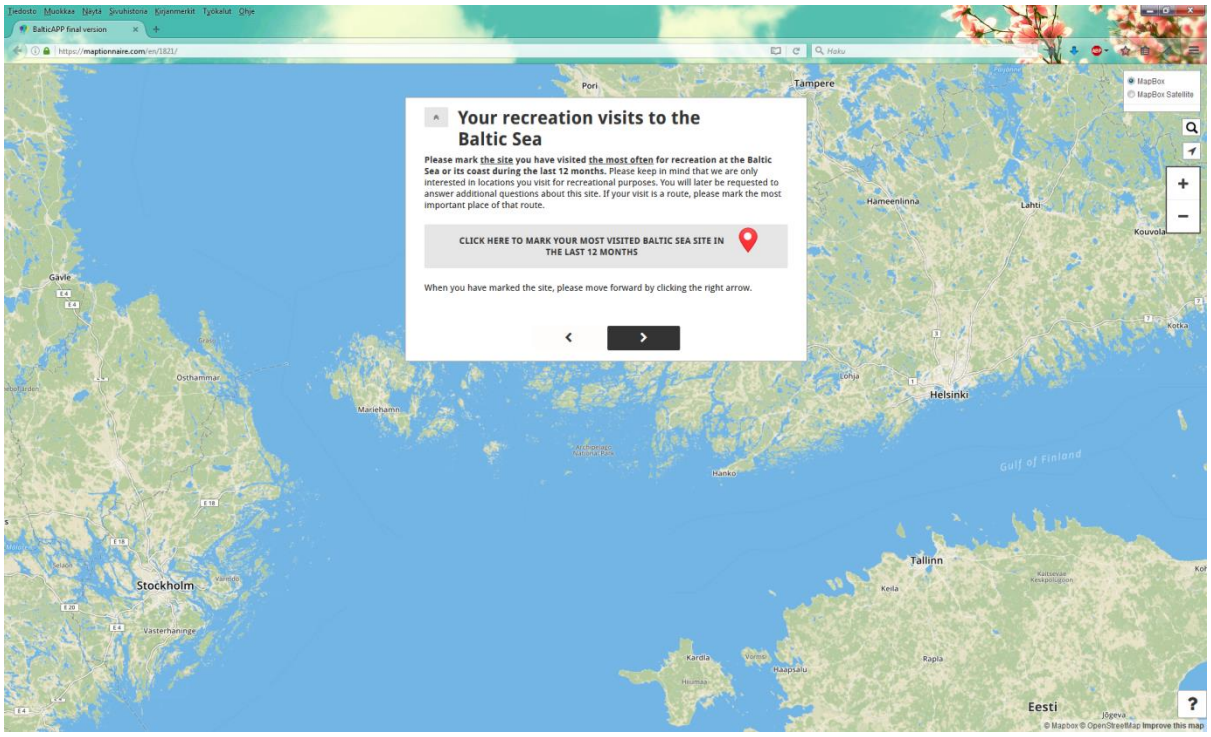
Q17d. A healthy ecosystem supports a large **diversity of native species**, including healthy populations of sea birds, plants and fish.

How would you describe the diversity of species at your most often visited site on average?

a) The number of bird and plant species is

Low	Rather low	Rather high	High	Don't know or remember
☐	☐	☐	☐	☐

Q20. Please think now that the environmental quality at the Baltic Sea and its coast becomes so poor that you start thinking about going somewhere else for recreation. Where in Country would you go to have a similar recreational experience? Please use again the map in the window below to mark this site. You can mark any site that you would consider going to instead of the Baltic Sea.



If you did not mark a location on the map, choose here:

- ○ I would stay at home
- ○ I don't know where I would go
- ○ Other reason

Q29. How important do you personally consider the following reasons to value the Baltic Sea and its coastal areas? Please consider your overall motivation to value the Baltic Sea as 100 points and distribute these points among the following reasons according to your motivations. It is not necessary to allocate points to all reasons, so if you do not care about the particular reason, you can write 0. The total must be 100 points.

Please distribute points	Points
Opportunities for recreational activities (e.g. swimming, fishing, walking, boating, bird watching).	____
Enjoyment from landscapes.	____
Inspiration for artistic work (photographing…).	____
An environment for learning and gaining new information.	____
Spiritual experiences, sense of belonging, and symbolic meaning.	____
Experiencing historically and culturally important places.	____
Habitats for many animals and plants.	____
Other reasons not mentioned in the list above.	____
Sum	100

Q36. What is your current occupational status? Please choose only one option that best describes your occupational status.

- ○ Employed full-time

- ○ Employed part-time
- ○ Retired
- ○ Student
- ○ Home-employed/Homemaker
- ○ Self-employed
- ○ Unemployed

Table: Variables for the Logistic regression

Dependent variable		Mean	Std. Dev.	Min	Max	n	p-value	Expected Sign
Substitution	0=No substitute site, 1= Substitute site	0,43	0,50	0	1	2642		
Independent variables								
Timing of visitation	0=Over 12 months, 1=Inside 12 months	0,76	0,43	0	1	2642	0,001	-/+
Trips to the Baltic Sea	0-365	12,12	41,04	0	365	2637	0,666	-/+
Summer visitation times divided with other seasons	0-1	0,56	0,36	0	1	2009	0,197	+
Hours stayed at most visited site	0,05-700 h	39,04	70,46	0	700	2621	0,001	-
Number of activities	0=one , 1=two or more	0,74	0,44	0	1	2544	0,000	+
Distance from respondents home to recreation site(stated)	0-1100 km	178,34	207,72	0	1100	2609	0,028	+
Euclidean distance from home to the Baltic Sea coast	0-689	83,35	108,23	0	689	2642	0,776	+
Euclidean distance from home to recreation site	0-1736	154,62	185,08	0	1736	2148	0,002	+
Trip duration (min)	0-1800	147,63	186,29	0	1800	2609	0,024	+
Importance of historical and cultural places	0-100 points	7,77	9,70	0	100	2641	0,264	-
Importance of recreation	0-100 points	33,20	23,64	0	100	2641	0,000	+
Importance of recreation (dummy)	0=0-25, 1=26-100	0,54	0,50	0	1	2641	0,000	+
Travel mode	a	3,24	1,11	1	7	2640	0,214	-/+
Travel mode dummy	0=Public Transport, 1=Car	0,79	0,41	0	1	2189	0,044	+
Travel mode dummy	0=Ferry, 1=Private boat	0,28	0,45	0	10	117	0,397	+
Separate activities								
Walking	0=No walking 1=Walking	0,68	0,47	0	1	2544	0,011	+
Jogging	0=No jogging, 1=Jogging	0,08	0,27	0	1	2544	0,048	+
Swimming	0=No swimming, 1=Swimming	0,32	0,47	0	1	2544	0,000	+
Angling	0=No angling, 1=Angling	0,05	0,23	0	1	2544	0,539	-/+
Nature watching	0=No nature watching, 1=Nature watching	0,31	0,46	0	1	2544	0,000	-/+
Diving	0=No diving, 1=Diving	0,01	0,07	0	1	2544	0,046	-
Boating	0=No boating, 1=Boating	0,07	0,25	0	1	2544	0,000	-/+
Sunbathing	0=No sunbathing, 1=Sunbathing	0,24	0,43	0	1	2544	0,004	+
Picnicking	0=No Picnicking, 1=Picnicking	0,20	0,40	0	1	2544	0,006	+

Summerhouse	0=No summerhouse, 1=Summerhouse	0,10	0,30	0	1	2544	0,966	-
Cruise	0=No cruise, 1=Cruise	0,11	0,32	0	1	2544	0,224	-
Sauna	0=No sauna, 1=Sauna	0,04	0,20	0	1	2544	0,014	-
Ice-skating	0=No Ice-skating, 1=Ice-skating	0,00	0,04	0	1	2544	0,769	-/+
Skiing	0=No skiing, 1=Skiing	0,01	0,08	0	1	2544	0,830	-/+
Conditions of the most visited site								
Location of recreation site	0=City, 20=Over 20 km from a city	6,03	9,18	0	20	2642	0,219	-/+
Location of recreation site	0=City, 30=Over 30 km from a city	10,07	14,17	0	30	2642	0,595	-/+
Location of recreation site	0=City, 40=Over 40 km from a city	14,88	19,34	0	40	2642	0,391	-/+
Location of recreation site	0=On continent or island, 1=Ocean	0,01	0,08	0	1	2642	0,526	-/+
Respondent perceived water clarity at most visited site	b	1,64	0,85	0	3	2374	0,753	-
Respondent perceived blue-algae	c	1,24	0,89	0	3	1814	0,302	+
Respondent perceived algae	c	1,47	0,83	0	3	2106	0,330	+
Respondent perceived number of species	d	1,64	0,73	0	3	2096	0,016	-
Respondent perceived number of species(dummy)	0=Low and rather low, 1=High and rather high	0,61	0,49	0	1	2096	0,036	-
Respondent perceived number of fish	d	1,37	0,78	0	3	936	0,467	-
Respondent perceived number of facilities	e	1,36	0,65	0	2	2497	0,000	-
Respondent perceived number of facilities(dummy)	0=None and some, 1=Many	0,45	0,50	0	1	2497	0,000	-
Demographics								
Age (categorical)	f	1,49	0,94	0	3	2462	0,000	-
Age (continuous)	18-79	46,13	15,37	18	79	2642	0,026	-
Age (dummy)	0=18-29 and over 65, 1=30-64	0,68	0,47	0	1	2642	0,000	-
Gender	0=male, 1=female	0,50	0,50	0	1	2642	0,633	-/+
Household size	1-7	2,32	1,12	1	7	2634	0,829	-
Under 18 year olds in a household	0-5	0,42	0,78	0	5	2638	0,214	-
Education (categorical)	g	2,82	1,03	1	4	2641	0,000	+

Education (dummy)	0=Other, 1=University	0,35	0,48	0	1	2641	0,000	+
Occupation (categorical)	h	2,51	1,84	1	7	2641	0,179	-/+
Occupation (dummy)	0=Other, 1=Fulltime	0,47	0,50	1	0	2641	0,729	-/+
Income (categorical)	i	4,15	2,08	1	8	2281	0,252	+
Income (continuous)	50-5000€	1812,90	1237,22	50	5000	2281	0,000	+
Country (categorical)	j	1,90	0,72	1	3	2642	0,000	-/+
Finland		0,48	0,50	0	1		0,000	-/+
Latvia		0,21	0,41	0	1		0,000	-/+
Germany		0,31	0,46	0	1		0,000	-/+

[a] Measured on a 7-point scale where 1=Walk, 2=Bike, 3=Car, 4=Public Transport, 5=Private Boat, 6=Ferry, 7=Other.
[b] Measured on a 4-point scale where 0=turbid, 1=somewhat turbid 2=somewhat clear, 3=clear.
[c] Measured on a 4-point scale where 0=never, 1=seldom, 2=sometimes, 3=often.
[d] Measured on a 4-point scale where 0=low, 1=rather low, 2=rather high, 3=high.
[e] Measured on a 3-point scale where 0=none, 1=some, 2=many.
[f] Measured on a 4-point scale where 0=18-29, 1=30-44, 2=45-64, 3=over 65.
[g] Measured on a 4-point scale where 1=Compulsory education, 2=Vocational education, 3=High School 4=University/Polytechnic.
[h] Measured on a 7-point scale where 1=Employed full-time, 2=Employed part-time, 3=Home-employed/Homemaker, 4=Retired, 5=Self-employed, 6=Student 7=Unemployed.
[i] Measured on a 8-point scale where 1=less than 200 euros, 2=201-500 euros, 3=501-900 euros, 4=901-1600 euros, 5=1601-2500 euros, 6=2501-3600 euros, 7=3601-5000 euros, 8=over 5000 euros.
[j] Measured on a 3-point scale where 1=Germany, 2=Finland, 3=Latvia.
[k] Measured on a 5-point scale where 1=Artificial surfaces, 2=Agricultural areas, 3=Forest and semi natural areas, 4=Wetlands, 5=Water bodies.
[l] Measured on a 5-point scale where 1=Artificial surfaces, 2=Agricultural areas, 3=Forest and semi natural areas, 4=Wetlands, 5=Water bodies.

Table: Probit model - selection equation

Probit model - selection equation		
	B	p-value
Constant	0,5672	0,000
Education	0,2094	0,000
Importance of recreation	0,2384	0,000
Age	-0,2264	0,001
Number of observations	3391	
Number of respondents	3395	
Log-likelihood	-1767,11	
Restricted Log-likelihood (constant only)	-1793,03	
Log-likelihood ratio	51,834	0,000
McFadden's pseudo R^2	0,0145	

Table: importance of recreation

	Importance of recreation				
	Mean	Std. Deviation	Min	Max	N
Substitute site	30,19	20,17	0	100	1133
No substitute site	35,46	25,72	0	100	1508